Audio-Visual Analysis In the Framework of Humans Interacting with Robots

Israel Dejene Gebru

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques et Informatique**

Arrêté ministériel :

Présentée par

Israel Dejene Gebru

Thèse dirigée par **Radu Horaud**

préparée au sein de **l'Inria Grenoble Rhône-Alpes**
et de **l'École Doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique**

Audio-Visual Analysis In the Framework of Humans Interacting with Robots

Thèse soutenue publiquement le **13 Avril 2018**,
devant le jury composé de :

Pr. Florence Forbes
INRIA Grenoble Rhône-Alpes, Présidente
Pr. Nicu Sebe
University of Trento (Italy), Rapporteur
Pr. Jean-Marc Odobez
IDIAP Research Institute (Switzerland), Rapporteur
Pr. Timothy Hospedales
University of Edinburgh (UK), Examinateur
Dr. Christine Evers
Imperial College London (UK), Invitée
Dr. Xavier Alameda-Pineda
INRIA Grenoble Rhône-Alpes, Co-Directeur de thèse
Pr. Radu Horaud
INRIA Grenoble Rhône-Alpes, Directeur de thèse

Audio-Visual Analysis In the Framework of Humans Interacting with Robots

Israel Dejene GEBRU

April 13, 2018

Abstract

In recent years, there has been a growing interest in human-robot interaction (HRI), with the aim to enable robots to naturally interact and communicate with humans. Natural interaction implies that robots not only need to understand speech and non-verbal communication cues such as body gesture, gaze, or facial expressions, but they also need to understand the dynamics of the social interplay, *e.g.,* find people in the environment, distinguish between different people, track them through the physical space, parse their actions and activity, estimate their engagement, identify who is speaking, who speaks to whom, etc. All these necessitate the robots to have multimodal perception skills to meaningfully detect and integrate information from their multiple sensory channels. In this thesis, we focus on the robot's audio-visual sensory inputs consisting of the (multiple) microphones and video cameras. Among the different addressable perception tasks, in this thesis we explore three, namely; (P1) multiple speakers localization, (P2) multiple-person location tracking, and (P3) speaker diarization. The majority of existing works in signal processing and computer vision address these problems by utilizing audio signals alone, or visual information only. However, in this thesis, we plan to address them via fusion of the audio and visual information gathered by two microphones and one video camera. Our goal is to exploit the complimentary nature of the audio and visual modalities with a hope of attaining significant improvements on robustness and performance over systems that use a single modality. Moreover, the three problems are addressed considering challenging HRI scenarios such as, *e.g.,* a robot engaged in a multi-party interaction with varying number of participants, which may speak at the same time as well as may move around the scene and turn their heads/faces towards the other participants rather than facing the robot.

We start this thesis by considering the problem of weighted data clustering. We propose a new mixture model that associates a weight with each observed point, namely, the Weighted-Data Gaussian Mixture Model (WD-GMM). We devise two EM algorithms to estimate model parameters. The first one, FW-EM, considers a fixed weight for each observation. The second one, WD-EM, treats each weight as a random variable following a gamma distribution. We provide a weight initialization strategy if the weight information is not available through a prior expert knowledge. Moreover, to determine the optimal number of mixture components, we introduce a model selection strategy based on Minimum Message Length (MML) criterion. We validate the proposed algorithms by comparing them with several state of the art parametric and non-parametric clustering techniques. The proposed mixture model and the EM algorithms form the basis for many of the works presented later in this thesis.

Subsequently we move on to the topic of audio-visual analysis, the core topic of this thesis, where we address the three problems mentioned above. Problem (P1) is addressed in the WD-GMM framework. We recast the speaker localization problem into audio-visual data clustering problem and present a methodology that optimally exploits the spatial coincidence of visual and auditory observations that are generated when people are both seen and heard. In particularly, we demonstrate a robust audio-visual speaker localization system, showing that such a system is more robust, more accurate and yields more

information than using a single modality. Moreover, through the application of WD-EM, we show that the pieces of information extracted from the audio and vision modalities are weighted accordingly to their relevance for speaker localization task to robustly discover coherent grouping of audio-visual observations; which represent the image locations of speaking persons. For Problem (P2), we propose a novel audio-visual tracking approach that exploits constructively the audio and visual modalities in order to estimate trajectories of multiple people in a joint state-space. The tracking problem is modeled using sequential Bayesian filtering framework, and density approximation techniques are used to keep the model tractable. The proposed tracking model is novel on how it represents the posterior distribution and the audio-visual likelihood with GMMs and how their form is kept the same over time, even though the number of Gaussian components would have been normally increasing. For Problem (P3), we propose a probabilistic diarization model by casting the problem into a speaker tracking formulation whereby the active speaker is detected and tracked over time. The proposed model uses on-image (spatial) coincidence of visual and auditory observations and sequentially infers a latent variable that represents the identity of the active speaker. The spatial coincidence is build on the intuition that a sound-source and associated visual-object should have the same spatial location. Consequently, it is possible to perform speaker localization by detecting and localizing persons in an image, estimating the directions of arrival of the active sound sources, mapping these sound directions onto the image and associating the dominant sound source with one of the persons that are visible in the image. All these is incorporated seamlessly in the proposed probabilistic model. In later work, we generalize this speaker diarization model for multiple speakers. We propose a speaker diarization model based on the fusion audio-visual data obtained from multiple-person visual tracking and multiple speakers (source) localization. We introduce a novel audio-visual association technique in order to tackle the speech-to-person association that often arises in multi-party dialogues. The diarization problem is casted into a latent-variable temporal graphical model that infers speaker identities and speech turns over time based on the output of an audio-visual association process executed at each time frame, and on the dynamics of the diarization variable itself. All four proposed models in this thesis were extensively evaluated using publicly available benchmarking datasets and a real-world audio-visual dataset which we collected and made publicly available. Our results and contributions have been peer-reviewed by the international research community and published in international top conferences and journals. We hope they will be essential elements of HRI systems in the future.

Résumé

Depuis quelques années, un intérêt grandissant pour les interactions homme-robot (HRI), avec pour but de développer des robots pouvant interagir (ou plus généralement communiquer) avec des personnes de manière naturelle. Cela requiert aux robots d'avoir la capacité non seulement de comprendre une conversation et signaux non verbaux associés à la communication (e.g. le regard et les expressions du visage), mais aussi la capacité de comprendre les dynamiques des interactions sociales, e.g. détecter et identifier les personnes présentes, où sont-elles, les suivre au cours de la conversation, savoir qui est le locuteur, à qui parle t-il, mais aussi qui regarde qui, etc. Tout cela nécessite aux robots d'avoir des capacités de perception multimodales pour détecter et intégrer de manière significative les informations provenant de leurs multiples canaux sensoriels. Dans cette thèse, nous nous concentrons sur les entrées sensorielles audio-visuelles du robot composées de microphones (multiples) et de caméras vidéo. Dans cette thèse nous nous concentrons sur trois tâches associés à la perception des robots, à savoir : (P1) localisation de plusieurs locuteurs, (P2) localisation et suivi de plusieurs personnes, et (P3) journalisation de locuteur. La majorité des travaux existants sur le traitement du signal et de la vision par ordinateur abordent ces problèmes en utilisant uniquement soit des signaux audio ou des informations visuelles. Cependant, dans cette thèse, nous prévoyons de les aborder à travers la fusion des informations audio et visuelles recueillies par deux microphones et une caméra vidéo. Notre objectif est d'exploiter la nature complémentaire des modalités auditive et visuelle dans l'espoir d'améliorer de manière significatives la robustesse et la performance par rapport aux systèmes utilisant une seule modalité. De plus, les trois problèmes sont abordés en considérant des scénarios d'interaction Homme-Robot difficiles comme, par exemple, un robot engagé dans une interaction avec un nombre variable de participants, qui peuvent parler en même temps et qui peuvent se déplacer autour de la scène et tourner la tête / faire face aux autres participants plutôt qu'au robot.

Nous commençons cette thèse en considérant le problème du regroupement de données pondérées. Nous proposons un nouveau modèle de mélange qui associe un poids à chaque point observé, à savoir le modèle de mélange gaussien de données pondérées (WD-GMM). Nous concevons deux algorithmes EM pour estimer les paramètres du modèle. Le premier, FW-EM, considère un poids fixe pour chaque observation. Le second, WD-EM, traite chaque poids comme une variable aléatoire suivant une distribution gamma. Nous fournissons une stratégie d'initialisation de poids si les informations de poids ne sont pas disponibles par le biais d'une expertise préalable. De plus, pour déterminer le nombre optimal de composants du mélange, nous introduisons une stratégie de sélection de modèle basée sur le critère de la longueur minimale du message (MML). Nous validons les algorithmes proposés en les comparant avec plusieurs techniques de regroupement paramétriques et non-paramétriques de l'état de l'art. Le modèle de mélange proposé et les algorithmes EM constituent la base de nombreux travaux présentés au cours de cette thèse.

Par la suite, nous abordons le thème de l'analyse audiovisuelle, thème central de cette thèse, où nous traitons les trois problèmes mentionnés ci-dessus. Le problème (P1) est traité dans le cadre de WD-GMM. Nous remanions le problème de localisation de lo-

cuteurs dans le problème de classification des données audiovisuelles et présentons une méthodologie qui exploite de façon optimale la coïncidence spatiale des observations visuelles et auditives générées lorsque les personnes sont à la fois vues et entendues. En particulier, nous démontrons qu'utiliser un système de localisation de locuteur à l'aide d'informations audio-visuel est plus robuste et plus précis et donne plus d'informations que l'utilisation d'une seule modalité. De plus, grâce à l'application de WD-EM, nous montrons que les informations extraites des modalités audio et vidéo sont pondérées en fonction de leur pertinence pour la localisation de locuteur afin de découvrir de manière robuste un regroupement cohérent d'observations audiovisuelles; qui représentent les emplacements visuels des locuteurs. Pour le problème (P2), nous proposons une nouvelle approche de suivi audiovisuel qui exploite de manière constructive les modalités auditives et visuelles afin d'estimer les trajectoires de plusieurs personnes dans un espace d'état commun. Le problème du suivi est modélisé en utilisant un cadre de filtrage bayésien séquentiel, et des techniques d'approximation de densité sont utilisées pour maintenir le modèle résolvable. Le modèle de suivi proposé est nouveau sur la façon dont il représente la distribution postérieure et la vraisemblance audiovisuelle avec les MGM et comment leur forme est maintenue dans le temps, même si le nombre de composants Gaussiens aurait normalement du augmenté. Pour le problème (P3), nous proposons un modèle de journalisation probabiliste en transformant le problème en une formulation de suivi de locuteur par laquelle le locuteur actif est détecté et suivi au cours du temps. Le modèle proposé utilise la coïncidence (spatiale) sur l'image des observations visuelles et auditives et déduit séquentiellement une variable latente qui représente l'identité du locuteur actif. La coïncidence spatiale repose sur l'intuition qu'une source sonore et un objet visuel associé doivent avoir le même emplacement spatial. Par conséquent, il est possible d'effectuer une localisation de locuteur en détectant des personnes dans une image, en estimant les directions d'arrivée des sources sonores actives, en mappant ces directions sonores sur l'image et en associant la source sonore dominante à l'une des personnes visibles dans l'image. Tous ces éléments sont intégrés de manière transparente dans le modèle probabiliste proposé. Dans un travail ultérieur, nous généralisons ce modèle de journalisation de locuteur pour plusieurs locuteurs. Nous proposons un modèle de journalisation de locuteur basé sur les données audiovisuelles de fusion obtenues à partir du suivi visuel de plusieurs personnes et de la localisation de multiples locuteurs. Nous introduisons une nouvelle technique d'association audiovisuelle pour aborder l'association de la parole à la personne qui apparaît souvent dans les dialogues multi parties. Le problème de journalisation est décomposé en un modèle graphique temporel avec des variables latentes qui infère les identités des locuteurs et les tours de parole dans le temps en fonction de la sortie d'un processus d'association audiovisuelle exécuté à chaque période et de la dynamique de la variable journalisation elle-même. Les quatre modèles proposés dans cette thèse ont été largement évalués à l'aide d'ensembles de données d'étalonnage disponibles au public et d'un ensemble de données audiovisuelles réelles que nous avons recueillies et rendues publiques. Nos résultats et contributions ont été évalués par la communauté internationale de publiés dans les meilleures conférences et revues internationales. Nous espérons qu'ils seront des éléments essentiels des systèmes d'interaction Homme-Robot dans le futur.

ACKNOWLEDGMENT

During the four years of my PhD, I had the chance to meet so many people who have contributed to this venture in so many ways, it would be impossible to list them all. I apologize for any omissions, and I hope you know that you are in my heart even if not on this page.

First and foremost, I would like to thank my advisor Dr. Radu Horaud for giving me the opportunity to work on a very interesting topic and for his invaluable guidance and support throughout my PhD. I thank him for supporting my research interests and providing me ample opportunities to independently explore and develop my ideas.

I'm very grateful to have Dr. Christine Ever, Pr. Florence Forbes, Pr. Jean-Marc Odobez, Pr. Nicu Sebe, Pr. Timothy Hospedales, Dr. Xavier Alameda-Pineda serving as my thesis committee members. I would like to thank you all for taking your time to read through my thesis and attend my PhD defense.

I had the opportunity to visit Imperial College London and collaborate with Christine Ever and Patrick A. Naylor. I would like to thank you for our very fruitful and efficient collaboration.

I take this opportunity to thank the wonderful people I met at Oculus Research in Pittsburgh, where I was fortunate enough to spend six months as a research intern. I had the chance to work with some of the smartest people I have ever met including Yaser Sheikh, Ravish Mehra, Dejan Markovic, Hatem Alismail, Mattew Stewart and Mohsen Shahmohamadi.

I would like to thank my colleagues and friends in Perception team and INRIA. I particularly want to thank Xavier, Sileye and Georgios for the inspiration and support they provided in the course of this research. I want to thank Vincent, Dionyssos, Yutong, Benoit, Stéphane, Laurent, Xiaofei, Soraya, Bastien, Guillaume, Fabien, Quentin, and so many other for their continuous support as well as making my time at INRIA very enjoyable. I think I have been exceptionally lucky to have been here with all of you. A very Special thanks to Nathalie for taking care of traveling and various administrative tasks.

My deepest gratitude to my family (my parents and my sisters) for their love, care and help, despite the long distances. Without their unconditional love and support, I could have neither started nor finished this journey.

Finally, I would like to thank Katarzyna: you always have been there for me. Thank you for your patience and unreserved support in this long process. Without your love none of this would have been possible. I look forward to many new adventures together.

"*Nobody tells this to people who are beginners, I wish someone told me. All of us who do creative work, we get into it because we have good taste. But there is this gap. For the first couple years you make stuff, it's just not that good. It's trying to be good, it has potential, but it's not. But your taste, the thing that got you into the game, is still killer. And your taste is why your work disappoints you. A lot of people never get past this phase, they quit. Most people I know who do interesting, creative work went through years of this. We know our work doesn't have this special thing that we want it to have. We all go through this. And if you are just starting out or you are still in this phase, you gotta know its normal and the most important thing you can do is do a lot of work. Put yourself on a deadline so that every week you will finish one story. It is only by going through a volume of work that you will close that gap, and your work will be as good as your ambitions. And I took longer to figure out how to do this than anyone I've ever met. It's gonna take a while. It's normal to take a while. You've just gotta fight your way through.*"

Ira Glass

CONTENTS

Acknowledgment **vii**

List of Figures **xvi**

List of Tables **xvii**

1 Introduction **1**

1.1 Motivations 1

1.2 Audio-Visual Analysis for HRI 2

1.3 Challenges in Audio-Visual Analysis for HRI 4

1.4 Audio-visual Fusion Strategies 4

1.5 Contributions of this Thesis 6

1.6 Organization of this Thesis 9

2 AV Dataset for Conversational Scene Analysis **11**

2.1 Introduction 11

2.2 Related Datasets 14

2.3 The AVDIAR Dataset 15

2.4 Recording Setup and Scenarios 16

2.4.1 Camera-Microphone Setup 16

2.4.2 Recorded Scenarios 17

2.5 Ground Truth Annotations 19

2.6 Audio-Visual Alignment 21

2.7 Conclusions 22

3 EM Algorithms for Weighted Data Clustering 23

3.1 Introduction . . . 23

3.2 Gaussian Mixture with Weighted Data . . . 26

3.3 EM with Fixed Weights . . . 27

3.3.1 The E-Step . . . 27

3.3.2 The M-Step . . . 28

3.4 Modeling the Weights . . . 28

3.5 EM with Random Weights . . . 29

3.5.1 The E-Z Step . . . 29

3.5.2 The E-W Step . . . 30

3.5.3 The Maximization Step . . . 31

3.6 Estimating the Number of Clusters . . . 32

3.7 Algorithm Initialization . . . 35

3.8 Benchmark Results . . . 35

3.9 Conclusions . . . 39

4 Weighted Data Clustering Applied to AV Speaker Localization 43

4.1 Introduction . . . 43

4.2 Related Work . . . 44

4.3 Audio-Visual Clustering . . . 46

4.3.1 The Audio Modality . . . 46

4.3.2 The Visual Modality . . . 47

4.3.3 Cross-modal Weighting . . . 48

4.3.4 Determining the Number of Speakers . . . 49

4.3.5 Post Processing . . . 49

4.4 Experiments . . . 50

4.4.1 Data Collection . . . 50

4.4.2 Results . . . 51

4.5 Conclusions . . . 53

5 AV Tracking by Density Approximation In A Sequential Bayesian Filtering Framework **55**

5.1 Introduction 55
5.2 System Model 57
5.3 Proposed Method 58
5.3.1 Predicted pdf 58
5.3.2 Likelihood 59
5.3.3 Posterior pdf 60
5.4 Experimental Setup 61
5.5 Experimental Validation 62
5.6 Conclusions 64

6 Tracking the Active Speaker based on AV Data **67**

6.1 Introduction 67
6.2 Model for Tracking the Active Speaker 69
6.2.1 The Audio-Visual Association Model 70
6.2.2 The State Transition Model 72
6.3 Implementation and Experiments 73
6.4 Conclusions 76

7 Speaker Diarization based on AV Spatiotemporal Fusion **79**

7.1 Introduction 79
7.2 Related Work 81
7.3 Proposed Model 83
7.3.1 Speaker Diarization Model 84
7.3.2 State Transition Model 86
7.4 Visual Observations 87
7.5 Audio Observations 88
7.5.1 Single Audio Source 88
7.5.2 Multiple Speech Sources 89
7.6 Audio-Visual Fusion 90
7.7 Experimental Evaluation 92

7.7.1 Audio-Visual Datasets 92
7.7.2 Diarization Algorithms and Setup 93
7.7.3 Diarization Performance Measure 95
7.7.4 Results and Discussion 95
7.8 Conclusions 97

8 Concluding and Future Directions **101**
8.1 Concluding Remarks 101
8.2 Direction for Future Research 104

Publications **105**
International Journal 105
International Conference Publications 105
Other Articles 105

References **107**

LIST OF FIGURES

2.1 A typical scenario in AVDIAR dataset 13

2.2 The **AVDIAR** dataset camera-microphone setup 17

2.3 Examples of scenarios in the **AVDIAR** dataset. In an effort to vary the acoustic conditions, we used three different rooms to record this dataset. . 19

3.1 Scatter plots of simulated data with and with outlier. 36

3.2 Results of fitting mixture models to the SIM-Mixed data in the presence outliers. 37

4.1 An example scenario in untethered human-robot-interaction with companion humanoid robot (NAO). 45

4.2 Pipeline for mapping binaural vector onto a sound-source direction. . . . 48

4.3 Sample frame from *fake speaker* (FS) sequence to demonstrate the application of WD-EM for speaker localization. 50

4.4 Speaker localization results on on the *fake speaker* (FS), *moving speaker* (MS) and *cocktail party* (CP) sequences. 52

5.1 Camera, microphone and recording room setup 61

5.2 Results obtained on scenario **S1** at video frame #351. 63

5.3 Results obtained on scenario **S1** at video frame #352. 63

5.4 Results obtained on scenario **S1** at video frame #451. 64

5.5 Results obtained on scenario **S1** at video frame #452. 64

6.1 Result obtained on the *counting* sequence. 75

6.2 Result obtained on the *chat* sequence. 75

6.3 Result obtained on the *Two10* sequence. 76

7.1 The Dynamic Bayesian Network (DBN) spatiotemporal fusion model for audio-visual speaker diarization. 85
7.2 Example frames from **MVAD**, **AVASM**, and **AV16P3** datasets. 94
7.3 Result obtained on sequence Seq32-4P-S1M1 from **AVDIAR** dataset. . . 98
7.4 Result obtained on sequence Seq12-3P-S2M1 from **AVDIAR** dataset. . . 99
7.5 Result obtained on sequence Seq12-3P-S2M1 from **AVDIAR** dataset. . . 99

LIST OF TABLES

2.1 Summary of recorded scenarios forming the **AVDIAR** dataset 20

3.1 Datasets used for benchmarking various clustering algorithms. 37

3.2 Benchmark clustering results on MNIST, WAV, BCW, and Letter Recognition datasets. 38

3.3 DB scores obtained on simulated datasets. 38

4.1 Speaker detection results comparasion. 53

5.1 Audio-Visual tracking results comparison. 62

7.1 List of sequences from the **AVDIAR** dataset used to evaluate the proposed diarization model. 93

7.2 DER scores obtained with MVAD dataset. 96

7.3 DER scores obtained with AVASM dataset. 96

7.4 DER scores obtained with AV16P3 dataset. 97

7.5 DER scores obtained with AVDIAR dataset. 98

CHAPTER 1

INTRODUCTION

This chapter provides an introduction to the main points of interest in this thesis. Section 1.1 introduces the motivation for audio-visual analysis in the context of **Human-Robot Interaction** (HRI), which is the primary topic of interest to this thesis. Section 1.2 describes a number of audio-visual analysis problems we aim to address in this thesis. Section 1.3 highlights some of the research challenges. Afterwards in Section 1.4 we describe audio-visual fusion approaches we followed to tackle some of the research challenges. Section 1.5 summarizes the contributions made during this thesis work. Section 1.6 provides an outline for the remainder of the thesis.

1.1 MOTIVATIONS

Rapid advances in technology and increasing expectations from machines have changed the face of robotics. In the past few years, robots have been predominately used in manufacturing and production environments to handle repetitive tasks that require very little (if any) interaction with humans. In recent years, however, there is an increasing interest to develop social robots that can work alongside humans and are friendly, communicative and socially interactive. These robots can thus be used at homes and in public spaces, *e.g.,* museums, hospitals, schools, etc., to educate, to entertain, or to assist humans. However, building robotic systems that naturally engage in social interactions with humans is extremely challenging. Natural interactivity implies humans communicating with robots in the same ways in which humans communicate with one another. And this specifically necessitate the robots to have a diverse set of complex skills; including mobility, communication, and perception abilities; all working in a coordinated fashion.

In particular, given that communication between humans is typically a multimodal process that simultaneously uses conversational speech and non-verbal communication cues (*e.g.,* body gesture, gaze and facial expression), it is crucial to endow robots with audio-visual perception abilities. In the last few years, vast investigations have been conducted by research fields such as artificial intelligence, signal processing and computer vision to bring a good level of audio-visual perception and understanding to robots using one

or more microphones and video cameras. Several algorithms and methods that facilitate natural modes of interaction have been developed. For example, using video cameras: to detect and track humans, recognize faces, body pose, gestures, actions and so on. Using microphones, *e.g.,* to localize and recognize the speaker, to recognize the speech, to catch important keywords, to diarize conversations, and more. However, due to the complexity of real HRI scenes, the use of video and audio as separate cues do not always provide optimum and robust solutions—since each modality necessarily has flaws or ambiguities. Moreover, there are certain perception tasks that could not be done easily, if at all, using unimodal analysis, *i.e.,* audio-only or visual-only approaches. For example, given only a single modality, and perhaps only a single sensor, disambiguating the audio from multiple speakers can be a challenge. An interesting and promising alternative is to combine the merits of the audio and visual modalities, which is also inspired by the fact that humans routinely perform tasks in which ambiguous auditory and visual data are combined in order to form a strong and accurate percept. Thus, in recent times the joint processing of audio and video data (henceforth, "Audio-Visual analysis") has been an increasing area of interest for researchers. The complementary characteristics of the two modalities can be exploit to overcome certain limitations and problems faced when working on a single modality.

1.2 AUDIO-VISUAL ANALYSIS FOR HRI

This thesis has at its core audio-visual analysis in the framework of humans interacting with robots. We are interested in analyzing the social interplay in situations where several people are present, *i.e.,* social human-robot interaction. For this, we use a configuration of audio-visual sensors consist of two microphones and a single video camera. Our main goal is to develop methods and algorithms for the robot that could provide audio-visual perception capabilities to achieve some level of natural interactivity with humans. Moreover, we plan to develop efficient techniques to combine/fuse/calibrate audio and video modalities in such a ways that one modality complement the weaknesses of the other modality. To that end, among the multitude of research problems, this thesis is concerned with three tasks, namely, (P1) Audio-Visual Speaker Localization, (P2) Audio-Visual Multi-Person Tracking, and (P3) Audio-Visual Speaker Diarization. In the following we briefly describe these tasks; including their uses specific to HRI and some of the challenges faced when working with unimodal approaches.

(P1) **Audio-Visual Speaker Localization**: this problem arises when the task is, *e.g.,* to detect people that are both seen and heard, such as active speaker(s) in the field of view (FoV) of the robot. This is a crucial skill for the robot in order to, *e.g.,* shift or focus its attention to a person that is likely to communicate with. Speaker localization could also provide rich spatial information to other auditory perception and signal processing tasks that seek to enhance a source signal by spatial-temporal filtering. However, because of the potentially large number of people moving and speaking in such cluttered environments the problem of robust speaker localization is challenging. For audio, the signal propagating from the speaker is usually corrupted by reverberation and multi-path effects and by background noise, making it difficult. For video, the camera view may be cluttered by

objects other than the speaker, and thus speaker faces may be totally or partially occluded, in which case face detection and localization is extremely unreliable. Moreover, the use of video alone to detect faces and then lip movements have to deal with the false detection of lip activity which is not easy. However, the use of audio and video cues jointly could alleviate some of the problems since sound and visual information are jointly generated when people speak, and they provide complementary advantages for speaker localization if their dependencies are jointly modeled. In this thesis, we plan to do that.

(P2) **Audio-Visual Multi-Person Tracking**: tracking in the simplest form can be defined as the problem of estimating the trajectories of people/object as they move around the scene. For robots knowledge of the locations of people is very crucial in order to accomplish any interactive tasks, and tracking can provide rich spatio-temporal information that can be useful to other perception tasks. For example, reliable face tracking is required to extract non-verbal cues (e.g. gaze and facial expressions). However, tracking multiple people is extremely challenging for a number of reasons. Often in scene that contains several people engaged in conversation, it is not feasible to disambiguate the audio from multiple speakers if only audio data is available. Moreover, natural periods of speech inactivity within sentences, prolonged inactivity during dialogues often leads to tracking failures due to difficulties to re-establish identities and continuous tracks. While detection and tracking is more suited to the visual modality, tracking is still challenging due to inadequate or unexpected changes in lights, difficulties to detect multiple people (or in general objects) in the scene and not to confuse similar-looking objects when they appear together, when they occlude each other, or when they are obscured by objects in the background.

(P3) **Audio-Visual Speaker Diarization**: this is the task of identifying "who spoke and when", or the assignment of speech signals to people engaged in a dialogue without any prior knowledge about the speakers involved nor their number. It is an important front-end component for many speech processing systems. For example, when two or more persons are engaged in a conversation, one important task to be solved prior to automatic speech recognition (ASR) and natural language processing (NLP), is to correctly assign speech segments to corresponding speakers. While many existing works have proposed methods that perform speaker diarization using audio-only, video-only or audio-visual data, they still do not address speaker diarization in on-line setting and in the presence of multiple simultaneous speakers. Most methods often assume the availability of long speech-turns with little (or no overlap); however, this assumption is violated in social HRI settings where speech-turns may last only a short-duration and several people may occasionally speak simultaneously. Moreover, audio-only methods have their difficulties since the signals received at the microphones are corrupted by environmental noise and reverberations. Video-only or audio-visual approaches that require the detection of frontal faces and of mouth/lip motions are also not feasible, since people may wander around, turn their heads away from the sensors, be occluded by other, or suddenly disappear from the robot FoV, and appear again later on.

1.3 CHALLENGES IN AUDIO-VISUAL ANALYSIS FOR HRI

While audio-visual analysis is likely to be more robust than audio-only or vision-only approaches, there are several challenges that need to be carefully addressed. The fusion of the two modalities introduces new challenges; in addition to the already challenging nature of "unimodal" analysis itself. The audio and visual modalities live in different spaces, they are contaminated by different noise types and levels with different distributions, and they are perturbed by different physical phenomena, *e.g.,* acoustic reverberations, lighting conditions, etc. Moreover, the two modalities have different spatio-temporal distributions. For example, a speaker may face the cameras/microphones while he/she is silent and may emit speech while he/she turns his/her face away. Speech signals have sparse spectro-temporal structure and they are mixed with other sound sources, such as music or background noise. Speaker faces may be totally or partially occluded, in which case face detection and localization is extremely unreliable. Thus, all these issues should have to be addressed to optimally fuse the two modalities in such ways that one modality compensate for weaknesses of the other.

The complexity of HRI scenes poses additional challenges. In typically untethered HRI interaction scenarios, humans are at some distance from the robot and in the scene observed by the robot's audio-visual sensors: (i) people may wonder around the scene, thus creating occlusions and limited field of observance, (ii) people may enter and leave the scene, thus the number of people changes over time, (iii) people may interact with others, thus they might not always face the robot's cameras and microphones, (iv) people may speak simultaneously and there might be other sound sources and background noise.

Moreover, robots have limited on-board computing resources. In order to enable seamless interaction, the robot's perception tasks must (re)act to humans actions and intents in (near) real-time. Therefore, audio-visual analysis models and algorithms for HRI applications should be computationally efficient, both in time and memory.

Additional challenging issue often neglected but need to be carefully address is audio-visual datasets. Standard audio-visual datasets are necessary for developing fusion strategies, training the audio-visual models and evaluating their performance. Such datasets are very difficult to collect, even harder to annotate and consequently not readily available. Also, given the broad scope of HRI, datasets collected for one specific task are not well suited for other tasks.

1.4 AUDIO-VISUAL FUSION STRATEGIES

The main objective of this thesis is to solve the aforementioned problems (P1, P2 and P3) by exploiting the complementary nature of audio and visual modalities in such way that we alleviate some of the challenges in unimodal analysis. This raises the question of how to efficiently combine the two modalities in different natural conditions and according to the task at hand. The first question to be addressed is where the fusion of the data should take place. There are several possibilities (refer to [Shivappa 10] for more detailed discussion). Some of the existing research works fuse information at the decision level, also called late-fusion. That is, audio and visual inputs are first processed by

modality-specific subsystems, whose outputs are subsequently combined, *e.g.,* by averaging individual scores. Another approach is feature-level fusion. In this case, audio and video features are concatenated into a larger feature vector which is then used as an input to further data processing. However, owing to the very different physical natures of audio and visual data, direct integration is not straightforward. There is no obvious way to associate dense visual maps with sparse sound sources [Khalidov 08a].

The approach that we follow in this thesis lies between these two extremes. The input features are first transformed into a common representation and the processing is then done based on the combined features in this representation. Within this strategy, we exploit the spatio-temporal coincidence of audio and visual observations generated by people when they are seen and/or heard. Spatial coincidence implies the audio and visual observations extracted at a given time, or through a short period of time corresponds to the same source (*e.g.,* a speaking person). The temporal coincidence implies observations from different modalities are grouped if their evolution is correlated through time. Spatio-temporal assumes the audio-visual observations coincidence occurs both in space and time. For example, the premise that a speech signal coincides with a person that is visible and that emits a sound.

The second question to be addressed is which features to select in order to best account for the individual and combined modalities and most importantly, each modality may compensate for weaknesses of the other one, especially in noisy conditions. In this thesis we choose to work on audio features that encode sound source location information. By the use of a microphone pair, certain characteristics, such as interaural time difference (ITD) and interaural level difference (ILD), can be computed as indicators of the position of the sound source [Deleforge 14c]. In the visual domain, we chose to use high-level visual features that are directly related to persons, *i.e.,* 2D image locations of faces, provided by either a face-detector, *e.g.,* using [Zhu 12] or a face-detector coupled with a visual tracker. However, a face-detector is only reliable when a frontal face is presented and uninformative if a person turned his/her face away from the camera. This has shown to be a limitation in many previous audio-visual analysis works, in which non-frontal detection was not possible. Instead, in this thesis, we first detect human upper-body [Bourdev 09, Ferrari 08] and then we approximate the bounding box that encloses the head. In this way we build a general-purpose visual person localizer that is robust to light changes and to pose, and that always provides a localization, refined in the case of frontal detection.

Another major challenge in audio-visual fusion is the representation of auditory and visual features. Since the two modalities features are in a different representation space (feature-space), we follow three different strategies to translate the auditory features to on-image observations, so as to lie in the same space as the visual features. However, all of them require a training data specific to the camera/microphone setup used. Therefore, for each camera/microphones setup used in this thesis, we collected audio-visual fusion training data (D1) using a loudspeaker marked with easily detectable visual marker. The loudspeaker is placed manually at different positions in the FoV of the camera/microphone setup and at each position a one second long white-noise is played and we simultaneously

recored the audio and the visual marker pixel position. The training data is a collection of white-noise sounds with associated pixel positions.

The first strategy (FS1) is based on the work in [Deleforge 14a]. It performs direction of arrival (DOA) estimation in 2D by mapping the spectral binaural cues (ITD and ILD) onto the image plane: a DOA estimate therefore corresponds to a pixel location in the image plane. The mapping is based on a regression model build by using the training data D1 to learn a probabilistic mapping from the source position space to the spectral domain. Moreover, the probabilistic framework provides the inversion mapping, thus a sound source localization mapping from spectral binaural cues. A prominent feature of the probabilistic model is the explicit modeling of missing data situations. That is to say, that the localization mapping does not need a spectral cue with meaningful information in all frequency bands. Instead, the mapping makes use of those frequency bands in which the source is emitting, and is able to decouple the content of the sound source from its position. The limitation of this method is it assume one a singe speaker is present, which is not always true for real scenarios. This strategy is used for AV speaker localization and AV active speaker tracking.

The second strategies (FS2) uses multiple-sound source localization algorithm followed by a geometric transformation to translate the DoAs to image pixel positions. The geometric transformation is inspired from the work in [Sanchez-Riera 12]. It uses a non-linear regression model trained on D1. We used this strategy to address the fusion problem we have faced in AV multi-person tracking task.

The third strategies (FS3) is based a lookup table with image pixel positions and spectral features as key-value pairs. The look-up table is built using audio-visual training data collected in a similar fashion as D1. The main objective is to build a dictionary so that for each pixel position in the training data, we can look for the corresponding spectral feature or vice-versa. For example, during testing time, given a query spectral features-vector extracted from the recorded audio signals, we can search the look-up table to get the spectral features that best matches the query vector, which in turn gives the pixel location corresponding to the direction of the sound sources. This strategy is used for AV speaker diarization.

1.5 CONTRIBUTIONS OF THIS THESIS

In the following, the main contributions of this thesis are outlined, and for each a brief description is provided.

AUDIO-VISUAL DATASET

Standard audio-visual datasets are necessary for developing fusion strategies, training the audio-visual models and evaluating their performance. In Chapter 2, we present the **AV-DIAR** dataset. This dataset is collected as part of the speaker diarization work proposed in Chapter 7 since existing AV datasets would not fulfill our requirements. First, most of them do not provide audio-visual training data we require to train our audio-visual fusion models. Second, we needed challenging multi-party conversational scenarios to

test the limits and failures of our models; however, most of the existing datasets for HRI tend to simplify the data and the environments, *e.g.,* a single person speaking at a time, people facing the robot, etc. Third, we needed audio-visual data acquired from a robot-centered view since our main focus is HRI. Therefore, we could not use datasets recorded in smart-rooms using spread camera/microphone networks. To satisfy all these requirements, we collected the **AVDIAR** dataset, and as a first contribution of this thesis, we made it publicly available to allow other researchers to evaluate the performances of their own audio-visual systems. The scenarios included in the dataset are that of multi-party conversations between a group of people which have been passively recorded using a stereo camera pair and multiple microphones attached on a dummy binaural head. Each scenario is annotated with bounding boxes of people faces and of upper-body positions, as well as their identity and speaking activity at each video frame. We believe the **AVDIAR** dataset will provide an excellent test-bed for researchers working on multimodal speaker localization, tracking and diarization.

EM ALGORITHMS FOR WEIGHTED DATA CLUSTERING

In Chapter 3, we address the problem of data clustering via Gaussian Mixture Models (GMM) which we enhance by introducing a data weighting scheme. This assumes for each data point in the dataset a weight value is provided to account for its reliability. The value may depend on expert or prior knowledge and may be experiment- or goal-dependent. Intuitively, higher the weight, more reliable the data point is. An interesting problem we would like to solve in this thesis is how do we perform data clustering incorporating such prior knowledge. Therefore, we introduce the weighted-data Gaussian mixture model (WD-GMM) and derived two EM algorithms. The first one considers a fixed weight for each data point. The second one treats each weight as a random variable. We propose a data-driven weight initialization scheme in the absence of any prior information/knowledge of the data weights. Additionally, we present a model selection strategy based on a Minimum Message Length (MML) criterion. Our experiments show that the WD-EM algorithm compares favorably with several state-of-the-art parametric and non-parametric clustering methods and it performs particularly well in the presence of a large number of outliers. This work has been used at different times in this thesis, *e.g.,* to perform robust audio-visual data clustering, to identify outliers, and to approximate a probability density from discrete samples. This work has been published in [Gebru 16a].

AV SPEAKER LOCALIZATION

In Chapter 4, we present audio-visual speaker localization algorithm. We address this on the premise that a speech signal coincides with a person that is visible and that emits a sound. This coincidence must occur both in space and time. We propose to group auditory features and visual features based on the premise that they share a common location if they are generated by the same speaker. Speaker related features are extracted from the raw audio and visual signals. The auditory features are then translated to on-image observations using a regression technique proposed in [Deleforge 14a], and hence they lie in the same feature-space as the visual observations. The locations of speaking person(s)

are obtained in a principled way through application of the WD-GMM EM algorithm to cluster the audio-visual observations. Note here that we introduced a cross-modal weighting scheme in which observations from one modality are weighted accordingly to their relevance, for the speaker localization task, by the other modality. And because of the cross-modal weights, the proposed algorithm is less affected by the presence of non-speaking faces and noisy auditory observations due to reverberations and background sounds. Our experiments show the proposed algorithm can easily disambiguate between speaking and non-speaking people in realistic scenarios with people speaking and moving around. This work at its first stage has been published in [Gebru 14] and a more solid version in [Gebru 16a].

AUDIO-VISUAL MULTI-PERSON TRACKING

In Chapter 5, we propose a novel audio-visual tracking approach that exploits constructively audio and visual modalities in order to estimate trajectories of multiple people in a joint state-space. Sound and visual information are jointly generated when people speak, and they provide complementary advantages for tracking if their dependencies are modeled jointly. Hence, multiple Direction-of-Arrival (DOA) estimates obtained by sound source localization and facial detections are used as audio-visual observations. The tracking problem is modeled using a sequential Bayesian filtering framework and density approximation techniques are used to keep the complexity tractable. The framework is novel on how it represents the posterior probability distribution and the audio-visual likelihood with Gaussian mixture models (GMM), and how it retain the same form sequentially over time, avoiding the exponential explosion of the number of Gaussians. This work has been published in [Gebru 17b] and got the Best Paper Award at HSCMA'17.

TRACKING THE ACTIVE SPEAKER BASED ON AV DATA

In Chapter 6, we deal with the problem of audio-visual active speaker tracking. The task is to detect and track the identity of the active speaker. We propose a probabilistic temporal graphical model that infers a single discrete latent variable that represents the identity of the active speaker sequentially over time. The model seamlessly combine the output a multi-person visual tracker and a sound-source localizer. The audio-visual observations likelihood is modeled as a generative model based on the WD-GMM, and it evaluates the probability of an observed person to be the active speaker, conditioned by the output of the multi-person visual tracker and the sound-source localizer. The transition prior which represents the speaking-turn dynamics is fulfilled by a multi-case transition model. Our experiments illustrate the proposed model effectiveness with challenging scenarios involving moving people who speak inside a reverberant room. This work has been published in [Gebru 15a, Gebru 15b].

AUDIO-VISUAL SPEAKER DIARIZATION

In Chapter 7, we address the problem of audio-visual speaker diarization. The model seamlessly combines auditory and visual data and is well suited for challenging scenarios that consist of several participants engaged in multi-party dialog, while they move

around and turn their head towards the other participants rather than facing the cameras and the microphones. Multiple-person visual tracking is combined with multiple speech source localization in order to tackle the speech-to-person association problem. The latter is solved within a novel audio-visual fusion method on the following grounds: binaural spectral features are first extracted from a microphone pair, then a supervised audio-visual alignment technique maps these features onto images, and finally a semi-supervised clustering method assigns binaural spectral features to visible persons. The main advantage of this method is that it processes in a principled way speech signals uttered simultaneously by multiple persons. The diarization is cast into a latent-variable temporal graphical model that infers speaker identities and speech turns, based on the output of the audio-visual association process available at each discrete time-step, and on the dynamics of the diarization variable itself. The proposed formulation yields an efficient exact inference procedure, and is thoroughly tested and benchmarked with respect to several state-of-the art diarization algorithms. This work has been published in [Gebru 17a].

1.6 Organization of this Thesis

This chapter has presented a brief introduction to this thesis work. The remainder of this thesis is organized in six chapters, each one presenting a contribution of this thesis as described in the previous section. Chapter 8 concludes the thesis and presents possible topics of future work in audio-visual analysis. The publications related to this thesis and the list of references are in backmatter. Enjoy your reading!

CHAPTER 2

AV DATASET FOR CONVERSATIONAL SCENE ANALYSIS

This chapter describes a novel dataset dedicated for audio-visual analysis of conversational social scenes, namely the **AVDIAR**(**A**udio-**V**isual **Diar**ization) dataset. Publicly available datasets in this field are limited to structured social settings such as round-table meetings and smart-room recordings, where audio-visual cues associated with seated participants can be reliably acquired through the use of a camera network and of a microphone array. The motivation behind the **AVDIAR** dataset is to enable audio-visual scene analysis of unstructured informal meetings and gatherings. For this purpose, we collected scenarios covering examples of interactions between a group of people engaged in informal conversations. The scenarios involve participants that are either static and speak or move and speak, in front of a camera-microphone unit placed on a dummy-head. In an attempt to record natural human-human interactions, all scenarios were unscripted and participants were allowed to enter, wander around, leave the scene at any time and interrupt each other while speaking. The recordings were performed in a standard reverberant environment. With these settings, we collected 23 sequences over a period of 15 days and with sequence duration ranging from ten seconds to three minutes (in total 27 minutes). We carefully annotated each frame with bounding boxes of faces and of upper-body positions, as well as their identity and speaking activity over the entire sequence duration. The acquisition setup, sequence annotation and content are fully detailed.

2.1 INTRODUCTION

The perception that we have about the world is influenced by elements of diverse nature. Indeed humans tend to integrate information coming from different sensory modalities to conveying/sharing information, understanding others' intentions/emotions and making decisions. Following this observation, over the past decades multimodal signal processing has been an increasing area of interest for researchers working in the field of signal processing and machine learning. In particular, with the availability of sensing devices such

as cameras and microphones, simultaneous analysis of audio and visual data has become an important part of this research because:

- it holds great potential overcoming certain limitations and problems faced when working on a single modality, *i.e.,* audio-only [Bregman 94] and video-only [Itti 98] approaches,
- it provides rich information that can be useful in a number of applications where understanding of people social behavior and interaction is essential, *e.g.,* in surveillance [Andersson 10], social signal processing (SSP) [Vinciarelli 12], Human-Romputer Interaction (HCI), HRI, etc.

Audio-visual analysis frameworks that address, for example, lower-level problems such as *speaker localization, identification, tracking, diarization* [Hershey 00, Cutler 00, Otsuka 07, Bohus 10] and high-level problems such as *extracting contextual and social interactions* [Zancanaro 06, Vinciarelli 12, Gatica-Perez 09] have been proposed. Promising results have been achieved in controlled settings *i.e.,* constrained and predictable scenarios; principally round-table meetings settings where audio and visual cues concerning orderly arranged participants can be reliably acquired through the use of network of distributed cameras and microphone array, often coupled with closed-view video cameras and close-talk microphones available for each participant [Zhou 08, Gatica-Perez 07]. However, the results are far from being satisfactory in unstructured social settings (*e.g.,* a situated multi-party interaction [Bohus 09, Johansson 14]) where people may enter and leave the scene at any time, may interact with the system and with others, and their goals, plans, and needs may change over time.

Interesting examples of unstructured social settings can be found in the context of HRI and HCI where robots (computers) are required to have conversational agents capable of interacting with people in open, unconstrained and crowded environments. For robots to fulfill such interactive tasks, they need to perform audio-visual scene analysis to understand the dynamic scene around them *e.g.,* detect and track people, identify speaking person, infer social cues, etc. However, unlike meeting room settings, in the HRI and HCI settings usually distant sensing devices are available; usually attached onto an artificial agent, *e.g.,* on a dummy-head as shown in Fig. 2.1. Thus, audio-visual scene analysis becomes extremely challenging problem in many perspectives. Due to visual clutters and occlusions, it is challenging to reliably detect humans and infer locations and head/body orientations. It is also challenging to distinguish between voice and other sounds, identifying the source of sound and recognize the speaker(s) from the mixed acoustic signals captured by the distant microphones.

Despite increasing need to develop socially aware robots (or machines) and proven advantage of multimodal data, there have been limited attempt to collect audio-visual data that can be used as a testbed to audio-visual research experiments, especially in unstructured social settings. Currently available datasets are somehow limited and designed for a very constrained and specific application, and tend to simplify the data and environment (see Section 2.2 for more details). Most of them use large sensor networks in which

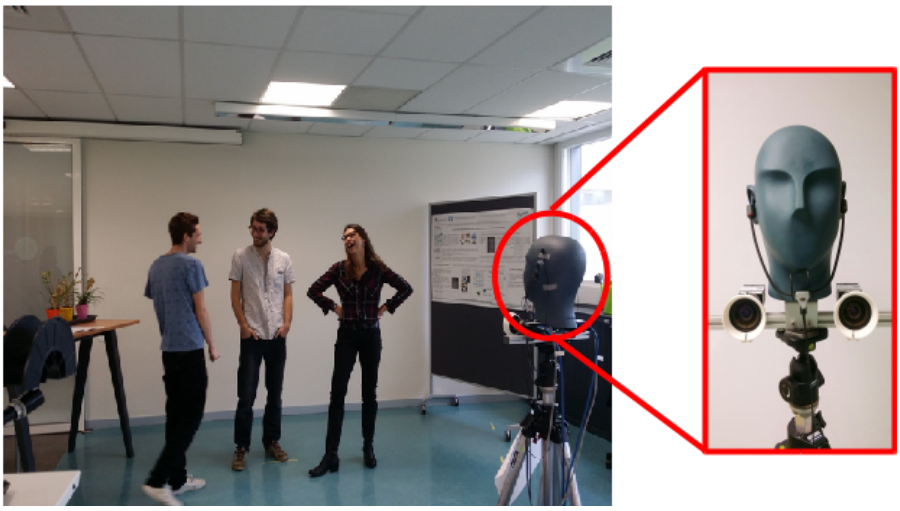


Figure 2.1: A typical scenario in AVDIAR dataset: multiple people engaged in a social conversation. A distant camera-microphone unit circled in red captures the scene globally.

microphones and cameras are often very close or even attached to people. Hence, they do not consider social and conversational scenario found in everyday real-world scenes. Moreover, most of them are provided with limited and sparse ground truth annotations.

To fill some of these research gaps, we proposed a new dataset, namely **AVDIAR** (**A**udio-**V**isual **Diar**ization) dataset. It covers examples of complex interactions between a group people engaged in a natural multi-party social conversation. The scenarios involve participants that are either static and speak or move and speak, in front of a camera-microphone unit placed on a dummy binaural head. In an attempt to record natural human-human interactions, the recordings were performed in a standard room with no special equipment and participants were allowed to wander around the scene and to interrupt each other while speaking. Furthermore, we carefully annotated each sequence in the dataset with participants' face and upper-body position as well as their identity and speaking activity over the entire sequence duration. Thus, the dataset is useful to assess the performance of audio-visual (or audio-only, video-only) analysis of methods using scenarios that were not available with existing datasets, *e.g.,* in a more realistic situation in which the participants were allowed to freely move in a room and to turn their heads towards the other participants, rather than always facing the camera.

The remainder of this chapter is organized as follows. Section 2.2 discusses a brief survey of audio-visual datasets and presents motivation for a new dataset. Section 2.3 introduces the **AVDIAR** dataset. Section 2.4 presents specifics of acquisition setup and details about the recorded scenarios. Section 2.5 describes ground truth annotation protocol and formats. Finally, Section 2.7 closes with some observations and suggestions about possible areas of research in audio-visual (or audio-only, video-only) processing that can be done using the **AVDIAR** dataset.

2.2 RELATED DATASETS

Numerous audio-visual datasets have been produced and publicly released in the two last decades. This section reviews (although not exhaustively) the existing audio-visual datasets and the most relevant ones are described.

Early research on audio-visual data started with audio-visual speech recognition (AVSR) and intelligibility research that focused on faces and speech. Many datasets were published in support *e.g.,* CUAVE [Patterson 02], AV-TIMIT [Hazen 04], GRID [Cooke 06], MOBIO [McCool 12] datasets, etc. These datasets include sequences with individual speaker or both individual speakers and speaker pairs acquired with a close-range fixed camera and a close-range microphone in a very quite environment. Some of these datasets were used to benchmark low-level tasks in audio-visual scene analysis *e.g.,* speaker localization [Gurban 06, Nock 03], person recognition [Zhao 12, Kächele 14]. However, a common weakness of many of the above dataset for audio-visual scene analysis is that the recorded video is highly constrained in terms of conversation flow and actions that subjects can take.

In the recent past there have been several research efforts to develop techniques and algorithms for automatic processing of multi-speaker meeting room data. To support the research efforts a number of multimodal meeting rooms (also called smart-rooms) equipped with multimodal sensors have been established by various research groups and consortia. Annotated audio-visual dataset have been collected and made publicly available, *e.g.,* the ISL [Burger 02], ICSI [Janin 03], NIST [Garofolo 04], AMI [Carletta 05], and AV16.3 [Lathoud 05]. ISL and ICSI datasets contain audio-only recording of natural meetings. The NIST dataset includes recordings from close-talking mics, lapel mics, distantly-placed mics, 5 video camera views, and full speaker/word-level transcripts. The AMI corpus contains recording from a wide range of devices including close-talking and far-field microphones, individual and room-view video cameras, projector, a whiteboard and individual pens. In the AV16.3 dataset three fixed cameras and two fixed 8-microphone circular arrays were used. These datasets involves multiple subjects who are involved in a continuous conversation with mostly one active speaker at particular time, which is true for a majority of the time in many meetings and presentations. Moreover, the NIST Rich Transcription (RT) evaluations in 2005 [Fiscus 05] and in 2006 [Fiscus 06], and CLEAR evaluations in 2006 [Stiefelhagen 06] and in 2007 [Stiefelhagen 08] have provided a common evaluation testbed to compare existing frameworks on specific tasks, *e.g.,* speaker detection, audio-visual (or audio-only, video-only) diarization, audio-visual tracking, etc.

One of the fundamental challenges in human-robot interactions (HRI) is providing robots with the audio-visual perception capabilities to interact with multiple human partners. Towards this goal many dataset have been published, *e.g.,* H3R [Mohammad 08], Vernissage corpus [Jayagopi 13], MHHRI [Celiktutan 17], etc. The H3R dataset contains audio-visual recording containing information of two of the modalities used by humans to communicate their emotional states, namely speech and facial expression. The Vernissage corpus provides a conversational multimodal dataset recorded with the humanoid robot NAO serving as an art guide introducing paintings to the participants and then quizzing

them in art and culture. A Wizard-of-Oz setup was used to manage the dialogue and control the robot's gaze and nodding. The interactions were recorded using three external cameras, two close-up micro- phones and a VICON motion capture system in addition to NAO's built-in camera and microphone, and were annotated with a set of nonverbal cues including speech utterances, 2D head location and visual focus of attention. The MHHRI dataset is introduce recently with the aim of studying personality simultaneously in human-human interactions (HHI) and HRI and its relationship with engagement. It is recorded using a set of sensors including two first-vision cameras (also known as egocentric cameras), two Kinect depth sensors and two physiological sensors.

Recently CAVA [Arnaud 08] and RAVEL [Alameda-Pineda 13] datasets have been published. They were specifically designed to model what the AV perception of the scene would be from a human point of view while limiting the recording equipment to a pair of binaural microphones and a pair of stereoscopic cameras. CAVA dataset contains audio-visual recordings collected using binocular and binaural camera-microphone pairs both mounted onto a person's. RAVEL dataset is recorded with 4 microphone and a stereoscopic camera fitted on a dummy head. These are two dataset closely related to **AVDIAR**.

Nevertheless, as described above different datasets used different devices to acquire the data, depending on the purpose. In the next section, the acquisition set up used in **AVDIAR**, which includes the recording environment and device, is fully detailed. Furthermore, the type of recorded data is specified as well as its main properties in terms of synchronization, calibration and audio-visual alignment.

2.3 THE AVDIAR DATASET

In order to provide a new and challenging evaluation framework for novel methodologies addressing challenges in audio-visual conversational scene analysis, we introduce the **AVDIAR** (Audio-Visual DIARrization) dataset. The **AVDIAR** dataset represents an excellent test-bed for research working on multimodal speaker localization, tracking and diarization of a group of people engaged in a complex conversational interactions due to the following reasons. First, a high quality audio-visual data is available. This include high resolution videos captured on a static stereoscopic camera setup with large field of view and a 6-channel audio data. Secondly, the scenarios were recorded from a first person perspective, thus the dataset could be an essential resource to develop tools and test algorithms for an efficient and relevant, robot-centered (limited by what the robot can sense), audio-visual analysis of conversational scenes. Thirdly, rich and dense ground-truth annotations are provided, at the video stream frame rate, *i.e.,* every 40ms for 25 FPS video. To the best of our knowledge, there is no such densely annotated audio-visual database publicly available. Finally, the scenarios were recorded in various indoor spaces and capture interaction of a group of people engaged in a natural social interplay. These three characteristics place **AVDIAR** in a unique position among the various audio-visual datasets available for detecting, localizing, and tracking of people engaged in a social conversation. The dataset has begun to be used: 14 recordings have already been successfully used to report results on audio-visual speaker diarization (see Chapter 7 and [Gebru 17a]).

A few sequences were also used in [Ban 17b] to test multi-speaker audio-visual tracking algorithm.

2.4 RECORDING SETUP AND SCENARIOS

Since the purpose of the **AVDIAR** dataset is to provide data for benchmarking methods and techniques working towards the understand of interactions by robots for the purpose of human-robot interaction, two requirements have to be addressed by the setup. First, a realistic and unconstrained recording environment is necessary. Second, a robo-centric viewpoint or first person perspective and hence the data is captured from the viewpoint of the observe. In this section the details of this setup are given, showing that the two requisites are satisfied to a large extent. In a first stage the recording device is described. Afterward, the acquisition environment is delineated. Finally, the properties of the acquired data in terms of quality and scenarios are detailed and discussed.

2.4.1 CAMERA-MICROPHONE SETUP

The camera-microphone setup consist of an acoustic dummy head, six microphones and two color cameras. A tripod is used to rigidly mount the acoustic dummy head and the two cameras, as shown in figure Fig. 2.1. The six microphones are then mounted on the dummy head, two of them are plugged into the left and right ears respectively, and the other four on each side of the head. The tripod height is adjusted at eye level of a normal height person, *i.e.,* at 1.5m above the ground.

The two microphones in the ears are from Sennheiser Triaxial MKE 2002 stereo microphone and the others four are Soundman OKM II Classic Solo microphones. They were all linked to the computer via the Behringer ADA8000 Ultragain Pro-8 digital external sound card sampling at 48kHz. Moreover, the usage of acoustic dummy head and the placement of microphones bring several advantages. First of all, since the head is alike of that of a human, it allows the possibility to carry out psycho-physical studies of sound *e.g.,* sound localization. Secondly, it allows for the comparison between using two microphones and Head Related Transfer Function (HRTF) against using four microphones with out HRTF. Finally, it fulfill the requirement mentioned in the beginning by delivering audio data captured from the viewpoint of the observer.

Each color camera is a PointGrey Grasshopper3 unit equipped with a Sony Pregius IMX174 CMOS sensor of size $1.2'' \times 1''$ and a Kowa 6 mm wide-angle lens. This camera-lens setup has a horizontal $\times$ vertical field of view of $97^\circ \times 80^\circ$, which is large enough for a range of HRI applications. The cameras deliver images with resolution of 1920×1200 color pixels at 25 FPS. They are placed roughly on a parallel plane with the ground and 20cm apart from each other, hence deliver stereoscopic images that are almost representative of a human-eye view of the scene. Most importantly, the stereoscopic setup allows to reconstruct the 3-D coordinates of an object from the 2-D coordinates in cameras' image planes. For this reason, stereo camera calibration is performed by the state-of-the art method described in [Bouguet 04] which uses several image-pairs to provide an accurate calibration.

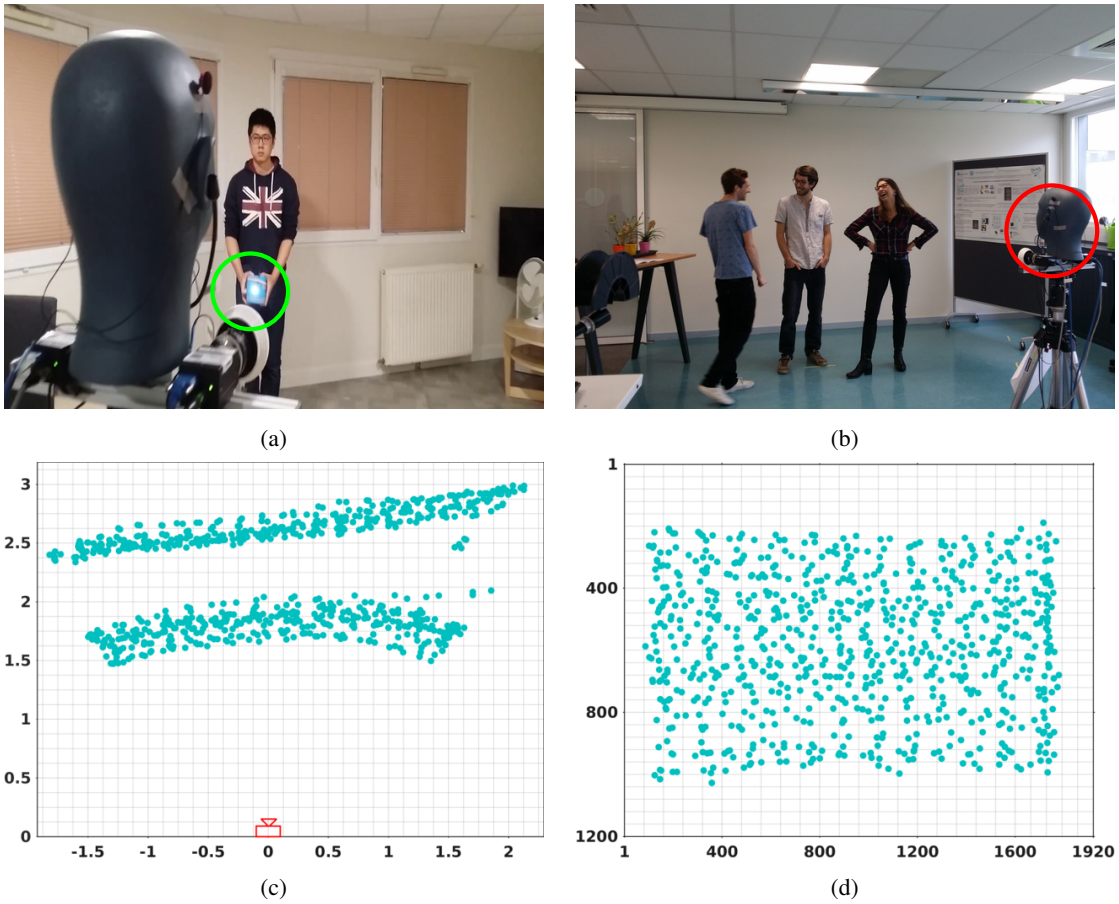


Figure 2.2: The **AVDIAR** dataset camera-microphone setup. (a) To record the training data, a loudspeaker that emits white-noise was used. A visual marker attached onto the loudspeaker (circled in green) allows to annotate the training data with image locations, each image location corresponds to a loudspeaker direction. (b) A typical **AVDIAR** scenario (the camera-microphone setup is circled in red). (c) Top-plane view of the loudspeaker position, the training data is collected roughly at depth of 1.5m and 2.5m from the camera. (d) The image grid of loudspeaker locations used for the training data.

Synchronized data, in terms of time, is increasingly important. The sound card and the two cameras are connected to a single PC and are finely synchronized by an external trigger controlled by a software. The audio-visual synchronization was done by using time-stamps delivered by the computer's internal clock with each audio and image frame.

2.4.2 Recorded Scenarios

The scenarios are recorded in a regular indoor space. For the sake of varying the acoustic and illumination condition in the recording, three different rooms are used: a standard living room, a regular meeting and an open office environment. They are shown in Fig. 2.3. Moreover, to keep a realistic recording environment, no effort was made to homogenize the environment by use of any backdrops nor to modify the acoustic properties of the rooms. Hence, the recordings are affected by exterior illumination changes, acoustic reverberations, outside noise, and all kind of audio and visual interferences and artifacts

present in unconstrained indoor scenes. For each recording room setup, the stereo camera and audio-visual calibration were performed as described in Section 2.4.1 and Section 2.6 respectively.

Each recorded scenario involves participants that are either static and speak or move and speak, in front of the camera-microphone unit at distance varying between 1.0m and 3.5m. In an attempt to record natural human-human interactions, participants were allowed to wonder around the scene and to interrupt each other while speaking. Moreover, none of participants were given any instructions on how to act in the form of a script. Consequently, the interaction dynamics correspond to those of a natural social interplay. We defined and recorded a set of sequences that contains a high variety of test cases: from short, very constrained, specific cases (e.g. visual occlusion), for each modality (audio or video), to natural spontaneous speech and/or motion in much less constrained context. We recorded the following scenario categories, *e.g.,* Fig. 2.3:

- *Static participants facing the camera.* This scenario can be used to benchmark algorithms and methods requiring the detection of frontal faces and of facial and lip movements.
- *Static participants facing each other.* This scenario can be used to benchmark algorithms and methods that require static participants not necessarily facing the camera.
- *Moving participants.* This is a general-purpose scenario that can be used to benchmark algorithms and methods such as speaker localization, audio-visual person tracking, diarization, etc.

The dataset comprises of 23 annotated sequences. It was recorded over a period of 15 days and with sequence duration ranging from ten seconds to three minutes (in total 27 minutes). Twelve different participants were recorded and upto five people are allowed in each sequence. Table 2.1 gives an overview of the recorded scenarios forming the **AVDIAR** dataset. In order to easy identify the sequence and its content a systematic name coding is used. The name of each sequence is unique and contains a compact description of its content. For example "SeqNN-**x**P-S**y**M**z**" has four parts:

- "SeqNN" is the unique identifier of this sequence.
- "**x**P" describes over all **x** number of different persons were recorded but not necessarily all visible at the same time.
- "S**y**", $\mathbf{y} \in \{0, 1, 2\}$ describes the auditory scene, 0 means there is no speech overlap, 1 means very little speech overlaps and 2 means often people speak at the same time and thus there are often long speech turn overlaps.
- "M**z**", $\mathbf{z} \in \{0, 1\}$ describes the visual scene, 0 means no or minor occlusion between participants and often the participants are static and face the camera and 1 means people wonder around the scene and often there are occlusions.

In addition to the real human-human interaction scenarios, we made a recording by placing a loudspeaker randomly at different position in the field of view of the camera. At each position, the loudspeaker emitted a 1 to 5 seconds random utterance from the TIMIT dataset [Garofolo 93]. The visual maker attached to the loudspeaker is used to extract the speech source direction, *e.g.,* Fig. **??**. We made such recording for each room setup. The duration of the each recording is around 22 minutes. The main purpose of these type of recordings is to test and benchmark various sound source localization algorithms.

Figure 2.3: Examples of scenarios in the **AVDIAR** dataset. In an effort to vary the acoustic conditions, we used three different rooms to record this dataset.

2.5 Ground Truth Annotations

Providing the ground truth is an important task when delivering a new data set; this allows to quantitatively compare the algorithms and techniques using the data. The **AVDIAR** dataset provides ground-truth annotations, which were performed either manually or semi-automatically over the entire sequence duration. For this specific purpose, a Matlab graphical interface program is written and used to facilitated the annotate task and create all annotation files available with the dataset. We present three type of annotations: visual, audio and audio-visual annotation. They are described as follows.

The visual annotation represents people locations and trajectories on the image sequences. The location of a person is describe in the form of a 2D bounding box on image plane. We annotated the head and upper-body bounding box of each person seen on the image. To describe the visual trajectory, each person is given a unique anonymous identity under the form of a digit $(0, 1, 2, \ldots)$ and this identity is consistent through the entire video sequence. The visual annotation is done in semi-automatic manner in three steps. First, a manual annotation of each person head (upper-body) bounding box is performed at a rate of one frame per second. Then, a visual tracker is employed to fill and interpolate the bounding box in the remaining frames. The main motivation to use a visual tracker is to save on time spent doing manual frame-by-frame annotations. Finally, manual frame-by-frame checking is done to correct tracker drifts and failures.

The audio annotation represents the speech/non-speech segmentation. The speech/non-speech segmentation is performed manually by marking the starting and ending time of

Table 2.1: Summary of recorded scenarios forming the **AVDIAR** dataset

Sequence name	Duration (seconds)	Description
Seq01-1P-S0M1	26.20	A single person moving randomly and alternating between speech and silence.
Seq02-1P-S0M1	28.00	
Seq03-1P-S0M1	23.00	
Seq04-1P-S0M1	18.64	
Seq37-2P-S0M0	43.08	Two static participants taking speech turns.
Seq43-2P-S0M0	16.36	
Seq05-2P-S1M0	34.84	Two static participants speaking almost simultaneously, *i.e.,* there are large speech overlaps.
Seq06-2P-S1M0	32.44	
Seq07-2P-S1M0	18.04	
Seq38-2P-S1M0	37.04	
Seq40-2P-S1M0	39.64	
Seq44-2P-S2M0	07.08	
Seq17-2P-S1M1	76.04	Two participants, wandering in the room and engaged in a conversation, sometime speaking simultaneously.
Seq18-2P-S1M1	76.04	
Seq19-2P-S1M1	68.04	
Seq20-2P-S1M1	54.04	
Seq21-2P-S1M1	45.04	
Seq08-3P-S1M1	103.00	Three participants engaged in an informal conversation. They are moving around and sometimes they speak simultaneously.
Seq09-3P-S1M1	65.00	
Seq10-3P-S1M1	19.00	
Seq12-3P-S1M1	78.84	
Seq13-4P-S2M1	144.04	Four participants engaged in an informal conversation. They are moving around, look in different directions and occasionally speak simultaneously. There are also frequent prolonged occlusions and interactions.
Seq32-4P-S1M1	39.44	
TIMIT-ROOM01	1320.00	Recordings made by placing a loudspeaker randomly at different position in the field of view of the camera. At each position, the loudspeaker emitted a random utterance from the TIMIT dataset [Garofolo 93].
TIMIT-ROOM02	1320.00	
TIMIT-ROOM02	1320.00	

speech utterances. However, it is very challenging to annotated very short utterances, *e.g.,* backchannel utterances such as “uh-huh", “oh", etc. We put special emphasis to annotate such short utterance whenever possible, *i.e.,* as long as they are clearly distinguishable in the audio mix and can be heard aloud. In general, we imposed a duration of 0.2 seconds

as a minimum speech duration and hence any utterance less than 0.2 second is simply ignored and automatically taken as a non-speech segment.

The audio-visual annotation represents the active speaker identities and locations. For each speech segments a set of target speaker identities is assigned referring back to the people identities defined as part of the visual annotation. The active speakers head bounding box centers are taken as the location of the speech sources.

To summarize, the three type of annotation provided with this dataset can be used to evaluate audio-visual speaker localization, tracking and diarization algorithms. Moreover, annotation is provided on each video frame, *i.e.,* it is very rich compared to other public available audio-visual datasets. The annotation files are provided in very popular file format as used in benchmark tools, *e.g.,* CSV files for visual annotations as used in MOTChallenge [1] and NIST RTTM files [2] for audio, and audio-visual annotations.

2.6 AUDIO-VISUAL ALIGNMENT

One very important issue with working with audio and visual data data representation. It is necessary to have an efficient data representation that is both appropriate for the task at hand and simplifies fusing the two modalities. For example, the most common choice in audio-visual tracking is to fuse the two modalities by first transforming them to a common space, *e.g.,* speech source directions and persons face directions. However this is not so easy since some form of microphone-camera calibration is needed. Recently, a few methods were proposed to address audio-visual data alignment, thus transforming auditory and visual data to a common representative space. A regression model is proposed in Deleforge et al. [Deleforge 14b] that map binaural features related to sound source direction with image pixel location. Khalidov et al. [Khalidov 13] propose to estimate the microphone locations into a camera-centered coordinate system and to use a binocular-binaural setup in order to jointly cluster visual and auditory feature using a mixture model [Khalidov 11a]. Gebru et al. [Gebru 17a] proposed to use a lookup table build with binaural features and image pixel location in audio-visual diarization task.

In order to satisfy the range of audio-visual alignment techniques used in the recent past and excite future line of research, we collect a training data consists of a recording of white-noise signals that are emitted with a loudspeaker and their directions. A visual marker, which can be easily detected and precisely located in an image, is placed onto the loudspeaker and allows to associate its image location with each sound direction. The recording is made by manually moving the loudspeaker in front of the camera-microphone unit, *e.g.,* Fig. 2.1. The loudspeaker is roughly moved in two planes roughly parallel to the image plane, at 1.5m and 2.5m, respectively. For each plane we record 800 positions lying on a uniform 20×40 grid that covers the entire field of view of the camera, hence there are 1600 training data samples.

[1] Benchmark for Multi-Target Tracking, `https://motchallenge.net/`

[2] Rich Transcription Spring 2006 Evaluation, `http://www.itl.nist.gov/iad/mig/tests/rt/2006-spring/`

2.7 CONCLUSIONS

This chapter introduces the **AVDIAR** dataset which consists of scenarios covering examples of natural interaction between two or more people engaged in informal conversations. In an attempt to record natural human-human interactions, all scenarios were unscripted and participants were allowed to enter, wander around, leave the scene at any time and interrupt each other while speaking. The acquisition setup, which consists of an acoustic dummy head, six microphones and two color cameras, was fully detailed. Technical specifications of the recorded streams (data) were provided. The dataset is carefully annotated each video frame with bounding boxes of faces and of upper-body positions, as well as their identity and speaking activity over the entire sequence duration. The scenarios were recorded from a first person perspective, thus the dataset could be an essential resource to develop tools and test algorithms for an efficient and relevant, robot-centered (limited by what the robot can sense), audio-visual analysis of conversational scenes.

CHAPTER 3

EM ALGORITHMS FOR WEIGHTED DATA CLUSTERING

Data clustering has received a lot of attention and numerous methods, algorithms and software packages are available. Among these techniques, parametric finite-mixture models play a central role due to their interesting mathematical properties and to the existence of maximum-likelihood estimators based on expectation-maximization (EM). This chapter proposes a new mixture model that associates a weight with each observed point. We introduce the weighted-data Gaussian mixture and we derive two EM algorithms. The first one considers a fixed weight for each observation. The second one treats each weight as a random variable following a gamma distribution. We propose a model selection (unsupervised model learning) method based on a Minimum Message Length (MML) criterion, provide a weight initialization strategy, and validate the proposed algorithms by comparing them with several state of the art parametric and non-parametric clustering techniques.

3.1 INTRODUCTION

Finding significant groups in a set of data points is a central problem in many fields. Consequently, clustering has received a lot of attention, and many methods, algorithms and software packages are available today. Among these techniques, parametric finite mixture models play a paramount role, due to their interesting mathematical properties as well as to the existence of maximum likelihood estimators based on expectation-maximization (EM) algorithms. While the finite Gaussian mixture (GMM) [McLachlan 00a] is the model of choice, it is extremely sensitive to the presence of outliers. Alternative robust models have been proposed in the statistical literature, such as mixtures of t-distributions [McLachlan 00b] and their numerous variants, e.g., [Bishop 05, Archambeau 07, Sun 10, Andrews 12, Forbes 14, Lee 14]. In contrast to the Gaussian case, no closed-form solution exists for the t-distribution and tractability is maintained via the use of EM and a

Gaussian scale mixture representation:

$$\mathcal{T}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha) = \int_0^\infty \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}/w)\ \mathcal{G}(w, \alpha/2, \alpha/2)dw, \tag{3.1}$$

where $\boldsymbol{x}$ is an observed vector, $\mathcal{N}$ is the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}/w$, and $\mathcal{G}$ is the gamma distribution of a univariate positive variable w parameterized by α. In the case of mixtures of t-distributions, with mixing coefficients π_k, $\sum_{k=1}^K \pi_k \mathcal{T}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_k)$, a latent variable w can also be introduced. Its distribution is a mixture of K gamma distributions that accounts for the component-dependent α_k [McLachlan 00b]. Clustering is then usually performed associating a positive variable w_i distributed as w with **each** observed point $\boldsymbol{x}_i$. Both the distributions of w_i and $\boldsymbol{x}_i$ do not depend on i. The observed data come from i.i.d. variables, distributed according to the t-mixture or one of its variants [McLachlan 00b, Bishop 05, Archambeau 07, Sun 10, Andrews 12, Forbes 14, Lee 14].

In this chapter we propose a finite mixture model in which variable w_i is used as a weight to account for the reliability of the observed $\boldsymbol{x}_i$ and this independently on its assigned cluster. The distribution of w_i is not a gamma mixture anymore but has to depend on i to allow each data point to be potentially treated differently. In contrast to mixtures of t-distributions, it follows that the observed data are independent *but not* identically distributed. We introduce the weighted-data Gaussian mixture model (WD-GMM). We distinguish two cases: (i) the weights are certainly known a priori and hence they are fixed, and (ii) the weights are modeled as random variables and hence they are iteratively updated, given initial estimates.

We show that in the case of fixed weights, the standard GMM parameters can be estimated via an extension of the standard EM which will be referred to as the *fixed weighted-data* EM algorithm (FWD-EM). Then we consider the more general case of weights that are treated as random variables. We model these variables with gamma distributions (one distribution for each variable) and we formally derive a closed-form EM algorithm which will be referred to as the *weighted-data* EM algorithm (WD-EM). While the M-step of the latter is similar to the M-step of FWD-EM, the E-step is considerably different as both the posterior probabilities (responsibilities) and the parameters of the posterior gamma distributions (the weights) are updated (E-Z-step and E-W-step). The responsibilities are computed using the Pearson type VII distribution (the reader is referred to [Sun 10] for a recent discussion regarding this distribution), also called the Arellano-Valle and Bolfarine generalized t-distribution [Kotz 04], and the parameter of the posterior gamma distributions are computed from the prior gamma parameters and from the Mahalanobis distance between the data and the mixture means.

Note that the weights play a different role than the responsibilities or the posteriori probabilities. Unlike the responsibilities, which are normalized, the weights are random variables that can take arbitrary positive values. Their posterior mean can be used as an absolute measure of the relevance of a datum. Typically, an outlying data point which is far from any cluster center will have a small weight while it may still be assigned with a significant responsibility value to the closest cluster. Responsibilities indicate which

cluster center is the closest but not if any of them is close at all.

The idea of weighted-data clustering has already been proposed in the framework of non-parametric clustering methods such as K-means and spectral clustering, e.g., [Long 06, Tseng 07, Ackerman 12, Feldman 12]. These methods generally propose to incorporate prior information in the clustering process in order to prohibit atypical data (outliers) from contaminating the clusters. The idea of modeling data weights as random variables and to estimate them via EM was proposed in [Forbes 10] in the particular framework of Markovian brain image segmentation. In [Forbes 10] it is shown that specific expert knowledge is not needed and that the data-weight distribution guide the model towards a satisfactory segmentation. A variational EM is proposed in [Forbes 10] as their formulation has no closed form. In this thesis we build on the idea that, instead of relying on prior information about atypical data, e.g., [Long 06, Tseng 07, Ackerman 12, Feldman 12], we devise a novel EM algorithm that updates the weight distributions. The proposed method belongs to the *robust clustering* category of mixture models because observed data that are far away from the cluster centers have little influence on the estimation of the means and covariances.

An important feature of mixture based clustering methods is to perform model selection on the premise that the number of components K in the mixture corresponds to the number of clusters in the data. Traditionally, model selection is performed by obtaining a set of candidate models for a range of values of K (assuming that the true value is in this range). The number of components is selected by minimizing a model selection criteria, such as the Bayesian inference criterion (BIC), minimum message length (MML), Akaike's information criteria (AIC) to cite just a few [McLachlan 00a, Figueiredo 02]. These methods have two main disadvantages. Firstly, a whole set of candidates has to be obtained and problems associated with running EM many times may emerge. Secondly, they provide a number of components that optimally approximates the density and not the true number of clusters present in the data. More recently, there seems to be a consensus among mixture model practitioners that a well-founded and computationally efficient model selection strategy is to start with a large number of components and to merge them [Hennig 10]. [Figueiredo 02] proposes a practical algorithm that starts with a very large number of components (thus making the algorithm robust to initialization), iteratively annihilates components, redistributes the observations to the other components, and terminates based on the MML criterion. [Baudry 10] starts with an overestimated number of components using BIC, and then merges them hierarchically according to an entropy criterion. More recently [Melnykov 14] proposes a similar method that merges components based on measuring their pair-wise overlap.

Another trend in handling the issue of finding the proper number of components is to consider Bayesian non-parametric mixture models. This allows the implementation of mixture models with an infinite number of components via the use of Dirichlet process mixture models. In [Rasmussen 99, Görür 10] an infinite Gaussian mixture (IGMM) is presented with a computationally intensive Markov Chain Monte Carlo implementation. More recently, more flexibility in the cluster shapes has been allowed by considering infinite mixture of infinite Gaussian mixtures (I^2GMM) [Yerebakan 14]. The flexibility is

however limited to cluster composed of sub-clusters of identical shapes and orientations, which may alter the performance of this approach. Altogether, IGMM and I^{2}GMM are not designed to handle outliers, as illustrated in Section 3.8, Figs. 3.2-f and 3.2-g. Infinite Student mixture models have also been considered [Wei 12], but inference requires a variational Bayes approximation which generates additional computational complexity.

Bayesian non-parametrics require a fully Bayesian setting. The latter, however, induces additional complexity for handling priors and hyper-priors, especially in a multivariate context. In contrast, our latent variable approach allows exact inference. With respect to model selection, we therefore propose to extend the method of [Figueiredo 02] to weighted-data mixtures. We formally derive an MML criterion for the weighted-data mixture model and we plug this criterion into an efficient algorithm which, starting with a large number of components, simultaneously estimates the model parameters, the posterior probabilities of the weights and the optimal number of components.

The remainder of this chapter is organized as follows. Section 3.2 outlines the weighted-data mixture model; Section 3.3 sketches the FWD-EM algorithm. Weights modeled with random variables are introduced in Section 3.4 and the WD-EM is described in detail in Section 3.5. Section 3.6 details how to deal with an unknown number of clusters and Section 3.7 addresses the issue of algorithm initialization. In Section 3.8 the proposed algorithms are tested and compared with several other parametric and non-parametric clustering methods. Section 3.9 concludes the chapter. Matlab code, additional results and videos are available on line[1].

3.2 Gaussian Mixture with Weighted Data

In this Section, we present the intuition and the formal definition of the proposed weighted-data model. Let $\boldsymbol{x} \in \mathbb{R}^d$ be a random vector following a multivariate Gaussian distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, namely $p(\boldsymbol{x}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, with the notation $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. Let $w > 0$ be a weight indicating the relevance of the observation $\boldsymbol{x}$. Intuitively, higher the weight w, stronger the impact of $\boldsymbol{x}$ should be. The weight can therefore be incorporated into the model by "*observing* $\boldsymbol{x}$ w *times*". In terms of the likelihood function, this is equivalent to raise $p(\boldsymbol{x}; \boldsymbol{\theta})$ to the power w, i.e., $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})^w$. However, the latter is not a probability distribution since it does not integrate to one. It is straightforward to notice that $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})^w \propto \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}/w)$. Therefore, w plays the role of the precision. Nevertheless, this model is not a standard Gaussian distribution because the weight w is different for each datum $\boldsymbol{x}$. Subsequently, we write:

$$\hat{p}(\boldsymbol{x}; \boldsymbol{\theta}, w) = \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}, \frac{1}{w}\boldsymbol{\Sigma}\right), \tag{3.2}$$

from which we derive a mixture model with K components:

$$\tilde{p}(\boldsymbol{x}; \boldsymbol{\Theta}, w) = \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_k, \frac{1}{w}\boldsymbol{\Sigma}_k\right), \tag{3.3}$$

[1] https://team.inria.fr/perception/research/wdgmm/

where $\boldsymbol{\Theta} = \{\pi_1, \ldots, \pi_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$ are the mixture parameters, $\pi_1, \ldots, \pi_K$ are the mixture coefficients satisfying $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$, $\boldsymbol{\theta}_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ are the parameters of the k-th component and K are the number of components. We will refer to the model in(3.3) as the *weighted-data Gaussian mixture model* (WD-GMM). Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ be the observed data and $\boldsymbol{W} = \{w_1, \ldots, w_n\}$ be the weights associated with $\boldsymbol{X}$. We assume each $\boldsymbol{x}_i$ is independently drawn from(3.3) with $w = w_i$. The observed-data log-likelihood writes:

$$\ln \tilde{p}(\boldsymbol{X}; \boldsymbol{\Theta}, \boldsymbol{W}) = \sum_{i=1}^{n} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \frac{1}{w_i}\boldsymbol{\Sigma}_k\right)\right). \tag{3.4}$$

It is well known that direct maximization of the log-likelihood function is problematic in case of mixtures and that the expected complete-data log-likelihood must be considered instead. Hence, we introduce a set of n hidden (assignment) variables $\boldsymbol{Z} = \{z_1, \ldots, z_n\}$ associated with the observed variables $\boldsymbol{X}$ and such that $z_i = k$, $k \in \{1, \ldots, K\}$ if and only if $\boldsymbol{x}_i$ is generated by the k-th component of the mixture. In the following we first consider a fixed (given) number of mixture components K, to later on extend the theory to unknown K, thus estimating the number of components from the data.

3.3 EM WITH FIXED WEIGHTS

The simplest case is when the weight values are provided at algorithm initialization, either using some prior knowledge or estimated from the observations (e.g., Section 3.7), and are then kept fixed while alternating between the expectation and maximization steps. In this case, the expected complete-data log-likelihood writes:

$$\mathcal{Q}_c\big(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(r)}\big) = \mathrm{E}_{P(\boldsymbol{Z}|\boldsymbol{X};\boldsymbol{W},\boldsymbol{\Theta}^{(r)})}\left[\ln P\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{W}, \boldsymbol{\Theta}\right)\right], \tag{3.5}$$

where $\mathrm{E}_P[\cdot]$ denotes the expectation with respect to the distribution P. The $(r+1)$-th EM iteration consists of two steps namely, the evaluation of the posterior distribution given the current model parameters $\boldsymbol{\Theta}^{(r)}$ and the weights $\boldsymbol{W}$ (E-step), and the maximization of(3.5) with respect to $\boldsymbol{\Theta}$ (M-step):

$$\boldsymbol{\Theta}^{(r+1)} = \arg\max_{\boldsymbol{\Theta}} \mathcal{Q}_c\big(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(r)}\big). \tag{3.6}$$

It is straightforward to show that this yields the following FWD-EM algorithm:

3.3.1 THE E-STEP

The posteriors $\eta_{ik}^{(r+1)} = p(z_i = k|\boldsymbol{x}_i; w_i, \boldsymbol{\Theta}^{(r)})$ are updated with:

$$\eta_{ik}^{(r+1)} = \frac{\pi_k^{(r)} \hat{p}(\boldsymbol{x}_i; \boldsymbol{\theta}_k^{(r)}, w_i)}{\tilde{p}(\boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, w_i)}, \tag{3.7}$$

where $\hat{p}$ and $\tilde{p}$ are defined in(3.2) and(3.3).

3.3.2 THE M-STEP

Expanding(3.5) we get:

$$\begin{aligned}\mathcal{Q}_c\big(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(r)}\big) &= \sum_{i=1}^{n}\sum_{k=1}^{K} \eta_{ik}^{(r+1)} \ln \pi_k \mathcal{N}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_k; \frac{1}{w_i}\boldsymbol{\Sigma}_k\right) \\ &\overset{\boldsymbol{\Theta}}{=} \sum_{i=1}^{n}\sum_{k=1}^{K} \eta_{ik}^{(r+1)} \Big(\ln \pi_k - \ln |\boldsymbol{\Sigma}_k|^{1/2} \\ &\quad - \frac{w_i}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\Big), \end{aligned} \tag{3.8}$$

where $\overset{\boldsymbol{\Theta}}{=}$ denotes equality up to a constant that does not depend on $\boldsymbol{\Theta}$. By canceling out the derivatives with respect to the model parameters, we obtain the following update formulae for the mixture proportions, means, and covariances matrices:

$$\pi_k^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n} \eta_{ik}^{(r+1)}, \tag{3.9}$$

$$\boldsymbol{\mu}_k^{(r+1)} = \frac{\sum_{i=1}^{n} w_i \eta_{ik}^{(r+1)} \boldsymbol{x}_i}{\sum_{i=1}^{n} w_i \eta_{ik}^{(r+1)}}, \tag{3.10}$$

$$\boldsymbol{\Sigma}_k^{(r+1)} = \frac{\sum_{i=1}^{n} w_i \eta_{ik}^{(r+1)} \left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(r+1)}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(r+1)}\right)^\top}{\sum_{i=1}^{n} \eta_{ik}^{(r+1)}}. \tag{3.11}$$

3.4 MODELING THE WEIGHTS

As we already remarked, the weights play the role of precisions. The notable difference between standard finite mixture models, and the proposed model is that there is a different weight w_i, hence a different precision, associated with *each* observation $\boldsymbol{x}_i$. Within a Bayesian formalism, the weights $\boldsymbol{W}$ may be treated as random variables, rather than being fixed in advance, as in the previous case. Since(3.2) is a Gaussian, a convenient choice for the prior on w, $p(w)$ is the conjugate prior of the precision with known mean, i.e., a gamma distribution. This ensures that the weight posteriors are gamma distributions as well. Summarizing we have:

$$P(w; \boldsymbol{\phi}) = \mathcal{G}(w; \alpha, \beta) = \Gamma^{-1}(\alpha)\, \beta^\alpha w^{\alpha-1} e^{-\beta w}, \tag{3.12}$$

where $\mathcal{G}(w; \alpha, \beta)$ is the gamma distribution, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\mathrm{d}t$ is the gamma function, and $\boldsymbol{\phi} = \{\alpha, \beta\}$ are the parameters of the prior distribution of w. The mean and

variance of the random variable w are given by:

$$\mathrm{E}[w] = \alpha/\beta, \tag{3.13}$$

$$\mathrm{var}[w] = \alpha/\beta^2. \tag{3.14}$$

3.5 EM WITH RANDOM WEIGHTS

In this section we derive the WD-EM algorithm associated to a model in which the weights are treated as random variables following(3.12). The gamma distribution of each w_i is assumed to be parameterized by $\phi_i = \{\alpha_i, \beta_i\}$. Within this framework, the expectation of the complete-data log-likelihood is computed over both the assignment and weight variables:

$$\mathcal{Q}_{\mathrm{R}}\big(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(r)}\big) = \mathrm{E}_{\mathrm{P}(\boldsymbol{Z},\boldsymbol{W}|\boldsymbol{X};\boldsymbol{\Theta}^{(r)},\boldsymbol{\Phi})}[\ln P(\boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{X}; \boldsymbol{\Theta}, \boldsymbol{\Phi})], \tag{3.15}$$

where we used the notation $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_n\}$. We notice that the posterior distribution factorizes on i:

$$P\big(\boldsymbol{Z}, \boldsymbol{W}|\boldsymbol{X}; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\Phi}\big) = \prod_{i=1}^{n} P\big(z_i, w_i|\boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i\big)$$

and each one of these factors can be be decomposed as:

$$\begin{aligned} &P\big(z_i, w_i|\boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i\big) = \\ &P(w_i|z_i, \boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i) P(z_i|\boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i), \end{aligned} \tag{3.16}$$

where the two quantities on the right-hand side of this equation have closed-form expressions. The computation of each one of these two expressions leads to two sequential steps, the E-W-step and the E-Z-step, of the expectation step of the proposed algorithm.

3.5.1 THE E-Z STEP

The marginal posterior distribution of z_i is obtained by integrating(3.16) over w_i. As previously, we denote the responsibilities with $\eta_{ik}^{(r+1)} = P(z_i = k|\boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i)$. The integration computes:

$$\begin{aligned} \eta_{ik}^{(r+1)} &= \int P\left(z_i = k, w_i|\boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i\right) dw_i \\ &\overset{k}{\propto} \int \pi_k^{(r)}\ P\left(\boldsymbol{x}_i|z_i = k, w_i; \boldsymbol{\Theta}^{(r)}\right)\ P(w_i; \boldsymbol{\phi}_i)\ dw_i \\ &= \int \pi_k^{(r)}\ \hat{p}\left(\boldsymbol{x}_i; \boldsymbol{\theta}_k^{(r)}, w_i\right)\ \mathcal{G}(w_i; \alpha_i, \beta_i)\ dw_i \\ &\propto \pi_k^{(r)}\ \mathcal{P}(\boldsymbol{x}_i; \boldsymbol{\mu}_k^{(r)}, \boldsymbol{\Sigma}_k^{(r)}, \alpha_i, \beta_i), \end{aligned} \tag{3.17}$$

where $\mathcal{P}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_i, \beta_i)$ denotes the Pearson type VII probability distribution function, which can be seen as a generalization of the t-distribution:

$$\mathcal{P}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \beta) = \frac{\Gamma(\alpha + d/2)}{|\boldsymbol{\Sigma}|^{1/2}\, \Gamma(\alpha)\, (2\pi\beta)^{d/2}} \left(1 + \frac{\|\boldsymbol{x} - \boldsymbol{\mu}\|^2_{\boldsymbol{\Sigma}}}{2\beta}\right)^{-(\alpha + \frac{d}{2})}. \tag{3.18}$$

3.5.2 THE E-W STEP

The posterior distribution of w_i, namely $p(w_i | z_i = k, \boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i)$ is a gamma distribution, because it is the conjugate prior of the precision of the Gaussian distribution. Therefore, we only need to compute the parameters of the posterior gamma distribution:

$$\begin{aligned} P(w_i | z_i = k, \boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i) &\overset{w_i}{\propto} \\ & P(\boldsymbol{x}_i | z_i = k, w_i; \boldsymbol{\Theta}^{(r)}) P(w_i; \boldsymbol{\phi}_i) \\ &= \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k^{(r)}, \boldsymbol{\Sigma}_k^{(r)}/w_i)\; \mathcal{G}(w_i; \alpha_i, \beta_i) \\ &= \mathcal{G}(w_i; a_i^{(r+1)}, b_{ik}^{(r+1)}), \end{aligned} \tag{3.19}$$

where the parameters of the posterior gamma distribution are evaluated with:

$$a_i^{(r+1)} = \alpha_i + \frac{d}{2}, \tag{3.20}$$

$$b_{ik}^{(r+1)} = \beta_i + \frac{1}{2} \left\| \boldsymbol{x}_i - \boldsymbol{\mu}_k^{(r)} \right\|^2_{\boldsymbol{\Sigma}_k^{(r)}} \tag{3.21}$$

The conditional mean of w_i, namely $\overline{w}_{ik}^{(r+1)}$, can then be evaluated with:

$$\overline{w}_{ik}^{(r+1)} = \mathrm{E}_{P(w_i | z_i = k, \boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i)}[w_i] = \frac{a_i^{(r+1)}}{b_{ik}^{(r+1)}}. \tag{3.22}$$

While estimating the weights themselves is not needed by the algorithm, it is useful to evaluate them in order to fully characterize the observations and to discriminate between inliers and outliers. First notice that the marginal posterior distribution of w_i is a mixture of gamma distributions:

$$\begin{aligned} & p(w_i | \boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i) \\ &= \sum_{k=1}^{K} p(w_i | z_i = k, \boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i) p(z_i = k | \boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i) \\ &= \sum_{k=1}^{K} \mathcal{G}(w_i; a_i^{(r+1)}, b_{ik}^{(r+1)})\; \eta_{ik}^{(r+1)}, \end{aligned} \tag{3.23}$$

and therefore the posterior mean of w_i is evaluated with:

$$\overline{w}_i^{(r+1)} = \mathrm{E}[w_i | \boldsymbol{x}_i; \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i] = \sum_{k=i}^{K} \eta_{ik}^{(r+1)} \overline{w}_{ik}^{(r+1)}. \tag{3.24}$$

By inspection of (3.20), (3.21), and (3.22) it is easily seen that the value of $\overline{w}_i$ decreases as the distance between the cluster centers and observation $\boldsymbol{x}_i$ increases. Importantly, the evaluation of $\overline{w}_i$ enables outlier detection. Indeed, an outlier is expected to be far from all the clusters, and therefore all $\overline{w}_{ik}$ will be small, leading to a small value of $\overline{w}_i$. It is worth noticing that this is not possible using only the responsibilities η_{ik}, since they are normalized by definition, and therefore their value is not an absolute measure of the datum's relevance, but only a relative measure of it.

3.5.3 The Maximization Step

This step maximizes the expected complete-data log-likelihood over the mixture parameters. By expanding(3.15), we have:

$$\begin{aligned}
&\mathcal{Q}_{\text{R}}\left(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(r)}\right) \\
&\quad \overset{\boldsymbol{\Theta}}{=} \sum_{i=1}^{n}\sum_{k=1}^{K}\int_{w_i} \eta_{ik}^{(r+1)} \ln \pi_k \mathcal{N}\Big(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \frac{1}{w_i}\boldsymbol{\Sigma}_k\Big) \times p(w_i|\boldsymbol{x}_i, z_i = k, \boldsymbol{\Theta}^{(r)}, \boldsymbol{\phi}_i)\mathrm{d}w_i \\
&\quad = \sum_{i=1}^{n}\sum_{k=1}^{K} \eta_{ik}^{(r+1)}\Big(\ln \pi_k - \ln|\boldsymbol{\Sigma}_k|^{1/2} - \frac{\overline{w}_{ik}^{(r+1)}}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\Big). \quad (3.25)
\end{aligned}$$

The parameter updates are obtained from canceling out the derivatives of the expected complete-data log-likelihood(3.25). As with standard Gaussian mixtures, all the updates are closed-form expressions:

$$\pi_k^{(r+1)} = \frac{1}{n}\sum_{i=1}^{n}\eta_{ik}^{(r+1)}, \quad (3.26)$$

$$\boldsymbol{\mu}_k^{(r+1)} = \frac{\sum_{i=1}^{n}\overline{w}_{ik}^{(r+1)}\eta_{ik}^{(r+1)}\boldsymbol{x}_i}{\sum_{i=1}^{n}\overline{w}_{ik}^{(r+1)}\eta_{ik}^{(r+1)}}, \quad (3.27)$$

$$\boldsymbol{\Sigma}_k^{(r+1)} = \frac{\sum_{i=1}^{n}\eta_{ik}^{(r+1)}\overline{w}_{ik}^{(r+1)}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(r+1)}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{(r+1)}\right)^\top}{\sum_{i=1}^{n}\eta_{ik}^{(r+1)}}. \quad (3.28)$$

It is worth noticing that the M-step of the WD-EM algorithm is strictly similar to the M-step of the FWD-EM algorithm (section 3.3). Indeed, the above iterative formulas, (3.26), (3.27), (3.28) are identical to the formulas (3.9), (3.10), (3.11), except that the fixed weights w_i are here replaced with the posterior means of the random weights, $\overline{w}_{ik}^{(r+1)}$.

3.6 ESTIMATING THE NUMBER OF CLUSTERS

So far it has been assumed that the number of mixture components K is provided in advance. This assumption is unrealistic for most real-world applications. In this section we propose to extend the method and algorithm proposed in [Figueiredo 02] to the weighted-data clustering model. An interesting feature of this model selection method is that it does not require parameter estimation for many different values of K, as it would be the case with the Bayesian information criterion (BIC) [Schwarz 78]. Instead, the algorithm starts with a large number of clusters and iteratively deletes clusters as they become irrelevant. Starting with a large number of clusters has the additional advantage of making the algorithm robust to initialization. Formally, the parameter estimation problem is cast into a transmission encoding problem and the criterion is to minimize the expected length of the message to be transmitted:

$$\text{length}(\boldsymbol{X}, \boldsymbol{\Theta}) = \text{length}(\boldsymbol{\Theta}) + \text{length}(\boldsymbol{X}|\boldsymbol{\Theta}). \tag{3.29}$$

In this context, the observations and the parameters have to be quantized to finite precision before the transmission. This quantization sets a trade off between the two terms of the previous equation. Indeed, when truncating to high precision, length($\boldsymbol{\Theta}$) may be long, but length($\boldsymbol{X}|\boldsymbol{\Theta}$) will be short, since the parameters fit well the data. Conversely, if the quantization is coarse, length($\boldsymbol{\Theta}$) may be short, but length($\boldsymbol{X}|\boldsymbol{\Theta}$) will be long. The optimal quantization step can be found by means of the Taylor approximation [Figueiredo 02]. In that case, the optimization problem corresponding to the *minimum message length* (MML) criterion, writes:

$$\begin{aligned}\boldsymbol{\Theta}_{\text{MML}} = \underset{\boldsymbol{\Theta}}{\text{argmin}} \Big\{ &- \log \text{P}(\boldsymbol{\Theta}) - \log \text{P}(\boldsymbol{X}|\boldsymbol{\Theta}, \boldsymbol{\Phi}) \\ &+ \frac{1}{2} \log|\mathbf{I}(\boldsymbol{\Theta})| + \frac{\mathcal{D}(\boldsymbol{\Theta})}{2}\Big(1 + \log \frac{1}{12}\Big)\Big\},\end{aligned} \tag{3.30}$$

where $\mathbf{I}(\boldsymbol{\Theta}) = -\text{E}\{D^2_{\boldsymbol{\Theta}} \log \text{P}(\boldsymbol{X}|\boldsymbol{\Theta})\}$ is the *expected* Fisher information matrix (FIM) and $\mathcal{D}(\boldsymbol{\Theta})$ denotes the dimensionality of the model, namely the dimension of the parameter vector $\boldsymbol{\Theta}$. Since the minimization (3.30) does not depend on the weight parameters, $\boldsymbol{\Phi}$ will be omitted for simplicity.

In our particular case, as in the general case of mixtures, the Fisher information matrix cannot be obtained analytically. Indeed, the direct optimization of the log-likelihood does not lead to closed-form solutions. Nevertheless, it was noticed that the *complete* FIM upper bounds the FIM [Figueiredo 02], and that the expected complete-data log-likelihood lower bounds the log-likelihood. This allows us to write the following equivalent optimization problem:

$$\begin{aligned}\boldsymbol{\Theta}_{\text{MML}} = \underset{\boldsymbol{\Theta}}{\text{argmin}} \Big\{ &- \log \text{P}(\boldsymbol{\Theta}) - \log \mathcal{Q}_{\text{R}}\big(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(r)}\big) \\ &+ \frac{1}{2} \log|\mathbf{I}_c(\boldsymbol{\Theta})| + \frac{\mathcal{D}(\boldsymbol{\Theta})}{2}\Big(1 + \log \frac{1}{12}\Big)\Big\},\end{aligned} \tag{3.31}$$

where $\mathbf{I}_c$ denotes the expected complete-FIM and $\mathcal{Q}_{\text{R}}$ is evaluated with (3.25).

As already mentioned, because there is a different weight w_i for each observation

i, the observed data are not identically distributed and our model does not account as a classical mixture model. For this reason, the algorithm proposed in [Figueiredo 02] cannot be applied directly to our model. Indeed, in the proposed WD-GMM setting, the complete-FIM writes:

$$\mathbf{I}_c(\boldsymbol{\Theta}) = \operatorname{diag}\Big(\pi_1 \sum_{i=1}^{n} \mathbf{I}_i(\boldsymbol{\theta}_1), \ldots, \pi_K \sum_{i=1}^{n} \mathbf{I}_i(\boldsymbol{\theta}_K), n\mathbf{M}\Big) \tag{3.32}$$

where $\mathbf{I}_i(\boldsymbol{\theta}_k) = -\mathrm{E}\{D^2_{\boldsymbol{\theta}_k} \log \mathcal{P}(\boldsymbol{x}_i|\boldsymbol{\theta}_k, \alpha_i, \beta_i)\}$ is the Fisher information matrix for the i-th observation with respect to the parameter vector $\boldsymbol{\theta}_k$ (mean and the covariance) of the k-th component, $\mathcal{P}$ is defined in (3.18), and $\mathbf{M}$ is the Fisher information matrix of the multinomial distribution, namely the diagonal matrix $\operatorname{diag}(\pi_1^{-1}, \ldots, \pi_K^{-1})$. We can evaluate $|\mathbf{I}_c(\boldsymbol{\Theta})|$ from (3.32):

$$|\mathbf{I}_c(\boldsymbol{\Theta})| = n^{K(M+1)} |\mathbf{M}| \prod_{k=1}^{K} \pi_k^M \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_i(\boldsymbol{\theta}_k) \right|, \tag{3.33}$$

where M denotes the number of free parameters of each component. For example, $M = 2d$ when using diagonal covariance matrices or $M = d(d+3)/2$ when using full covariance matrices.

Importantly, one of the main advantages of the methodology proposed in [Figueiredo 02] is that one has complete freedom to choose a prior distribution on the parameters, $\mathrm{P}(\boldsymbol{\Theta})$. In our case, inspired by (3.33), we select the following prior distributions for the parameters:

$$\mathrm{P}(\boldsymbol{\theta}_k) \propto \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{I}_i(\boldsymbol{\theta}_k) \right|^{-\frac{1}{2}}, \tag{3.34}$$

$$\mathrm{P}(\pi_1, \ldots, \pi_K) \propto |\mathbf{M}|^{-\frac{1}{2}}. \tag{3.35}$$

By substitution of (3.33)–(3.35) into (3.31) we obtain the following optimization problem:

$$\begin{aligned} \boldsymbol{\Theta}_{\text{MML}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \Big\{ & \frac{M}{2} \sum_{k=1}^{K} \log \pi_k - \log \mathcal{Q}_{\text{R}}\left(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(r)}\right) \\ & + \frac{K(M+1)}{2} \left(1 + \log \frac{n}{12}\right) \Big\}, \end{aligned} \tag{3.36}$$

where we used $\mathcal{D}(\boldsymbol{\Theta}) = K(M+1)$.

One may notice that (3.36) does not make sense (diverges) if any of the π_k's is allowed to be null. However, in the current message length coding framework, there is no point in transmitting the parameters of an empty component. Therefore, we only focus on the non-empty components, namely those components for which $\pi_k > 0$. Let $\mathcal{K}^+$ denote the index set of non-empty components and let $K^+ = |\mathcal{K}^+|$ be its cardinal. We can rewrite

(3.36) as:

$$\boldsymbol{\Theta}_{\text{MML}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \Big\{ \frac{M}{2} \sum_{k \in \mathcal{K}^+} \log \pi_k - \log \mathcal{Q}_{\text{R}} \left(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(r)} \right) + \frac{K^+(M+1)}{2} \left(1 + \log \frac{n}{12} \right) \Big\}. \tag{3.37}$$

The above minimization problem can be solved by modifying the EM algorithm described in Section 3.5 (notice that there is an equivalent derivation for the fixed-weigth EM algorithm described in Section 3.3). Indeed, we remark that the minimization (3.37) is equivalent to using a symmetric improper Dirichlet prior for the proportions with exponent $-M/2$. Moreover, since the optimization function for the parameters of the Gaussian components is the same (equivalently, we used a flat prior for the mean vector and covariance matrix), their estimation formulas (3.27) and (3.28) still hold. Therefore, we only need to modify the estimation of the mixture proportions, namely:

$$\pi_k = \frac{\max\left\{0, \sum_{i=1}^n \eta_{ik} - \frac{M}{2}\right\}}{\sum_{k'=1}^K \max\left\{0, \sum_{i=1}^n \eta_{ik'} - \frac{M}{2}\right\}}. \tag{3.38}$$

The $\max$ operator in (3.38) verifies whether the k-th component is supported by the data. When one of the components becomes too weak, i.e., the required minimum support $M/2$ cannot be obtained from the data, this component is annihilated. In other words, its parameters will not be estimated, since there is no need in transmitting them. One has to be careful in this context, since starting with a large value of K may lead to several empty components. In order to avoid this singular situation, we adopt the component-wise EM procedure (CEM) [Celeux 01], as proposed in [Figueiredo 02] as well. Intuitively, we run both E and M steps for one component, before moving to the next component. More precisely, after running the E-Z and E-W steps for the component k, its parameters are updated if $k \in \mathcal{K}^+$, otherwise the component is annihilated if $k \notin \mathcal{K}^+$. The rationale behind this procedure is that, when a component is annihilated its probability mass is immediately redistributed among the remaining components. Summarizing, CEM updates the components one by one, whereas the classical EM simultaneously updates all the components.

The proposed algorithm is outlined in Algorithm 1. In practice, an upper and a lower number of components, K_{high} and K_{low}, are provided. Each iteration r of the algorithm consists in component-wise E and M steps. If needed, some of the components are annihilated, and the parameters are updated accordingly, until the relative length difference is below a threshold, $\left|\Delta \mathbf{LEN}^{(r)}_{\text{MML}}\right| < \varepsilon$. In that case, if the message length, i.e., (3.37) is lower than the current optimum, the parameters, weights, and length are saved in $\boldsymbol{\Theta}_{\min}$, $W_{\min}$ and $\mathbf{LEN}_{\min}$ respectively. In order to explore the full range of K, the less populated component is artificially annihilated, and CEM is run again.

3.7 ALGORITHM INITIALIZATION

The EM algorithms proposed in Section 3.3, Section 3.5, and Section 3.6 require proper initialization of both the weights (one for each observation and either a fixed value w_i or parameters α_i, β_i) and of the model parameters. The K-means algorithm is used for an initial clustering, from which values for the model parameters are computed. In this section we concentrate onto the issue of weight initialization. An interesting feature of our method is that the only constraint on the weights is that they must be positive. Initial w_i values may depend on expert or prior knowledge and may be experiment- or goal-dependent. This model flexibility allows the incorporation of such prior knowledge. In the absence of any prior information/knowledge, we propose a data-driven initialization scheme and make the assumption that densely sampled regions are more important that sparsely sampled ones. We note that a similar strategy could be used if one wants to reduce the importance of dense data and to give more importance to small groups of data or to sparse data.

We adopt a well known data similarity measure based on the Gaussian kernel, and it follows that the weight w_i associated with the data point i is evaluated with:

$$w_i = \sum_{j \in \mathcal{S}_i^q} \exp\left(-\frac{d^2(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\sigma}\right), \tag{3.39}$$

where $d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the Euclidean distance, $\mathcal{S}_i^q$ denotes the set containing the q nearest neighbors of $\boldsymbol{x}_i$, and σ is a positive scalar. In all the experiments we used $q = 20$ for the simulated datasets and $q = 50$ for the real datasets. In both cases, we used $\sigma = 100$. In the case of the FWD-EM algorithm, the weights w_i thus initialized remain unchanged. However, in the case of the WD-EM algorithm, the weights are modeled as latent random variables drawn from a gamma distribution, hence one needs to set initial values for the parameters of this distribution, namely α_i and β_i in (3.12). Using (3.13) and (3.14) one can choose to initialize these parameters such as $\alpha_i = w_i^2$ and $\beta_i = w_i$, such that the mean and variance of the prior distribution are w_i and 1 respectively.

3.8 BENCHMARK RESULTS

The proposed algorithms were tested and evaluated using eight datasets: four simulated datasets and four publicly available datasets that are widely used for benchmarking clustering methods. The main characteristics of these datasets are summarized in Table 3.1. The simulated datasets (SIM) are designed to evaluate the robustness of the proposed method with respect to outliers. The simulated inliers are drawn from Gaussian mixtures while the simulated outliers are drawn from a uniform distribution, e.g., Fig. 3.1. The SIM datasets have different cluster configurations in terms of separability, shape and compactness. The eight datasets that we used are the following:

- **SIM-Easy:** Five clusters that are well separated and compact.
- **SIM-Unbalanced:** Four clusters of different size and density.

- **SIM-Overlapped:** Four clusters, two of them overlap.
- **SIM-Mixed:** Six clusters of different size, compactness and shape.
- **MNIST** contains instances of handwritten digit images normalized to the same size [LeCun 98]. We preprocessed these data with PCA to reduce the dimension from 784 to 141, by keeping 95% of the variance.
- **Wav** is the Waveform Database Generator [Breiman 84].
- **BCW** refers to the Breast Cancer Wisconsin data set [Street 93], in which each instance represents a digitized image of a fine needle aspirate (FNA) of breast mass.
- **Letter Recognition** contains $20,000$ single-letter images that were generated by randomly distorting the images of the 26 uppercase letters from 20 different commercial fonts [Frey 91]. Each letter/image is described by 16 features. This dataset is available through the UCI machine learning repository.

In addition to the two proposed methods (FWD-EM and WD-EM) we tested the following algorithms:

- **GMM** uses EM with the standard Gaussian mixture model, implemented as described in [Bishop 06];
- **GMM+U** uses EM with a GMM and with an additional uniform component, [Banfield 93];
- **FM-uMST** stands for the *finite mixture of unrestricted multivariate skew t-distribution* algorithm of [Lee 14];
- **IGMM** stands for the *infinite Gaussian mixture model* [Rasmussen 99];
- **I^{2}GMM** stands for the *infinite mixture of infinite Gaussian mixtures* [Yerebakan 14];

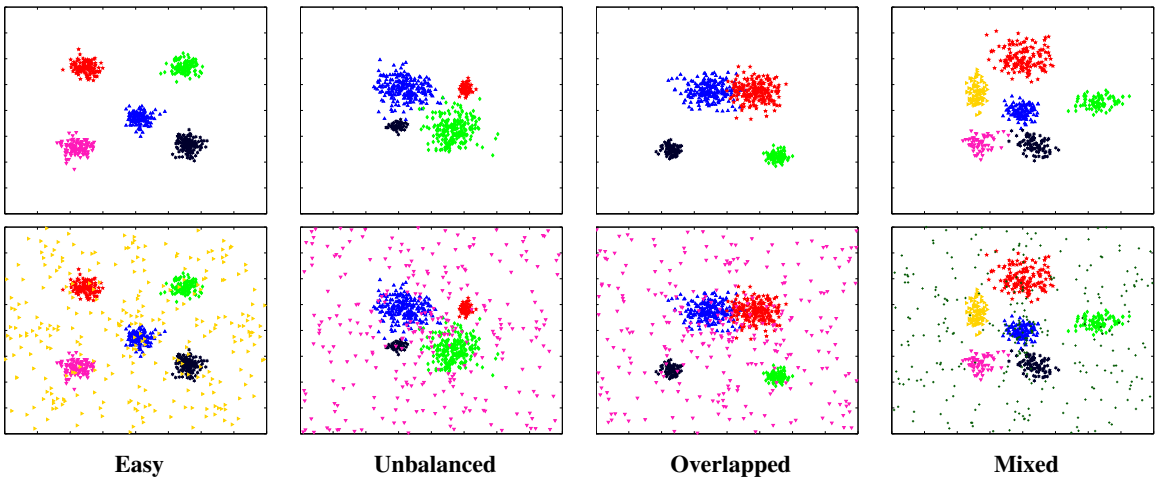


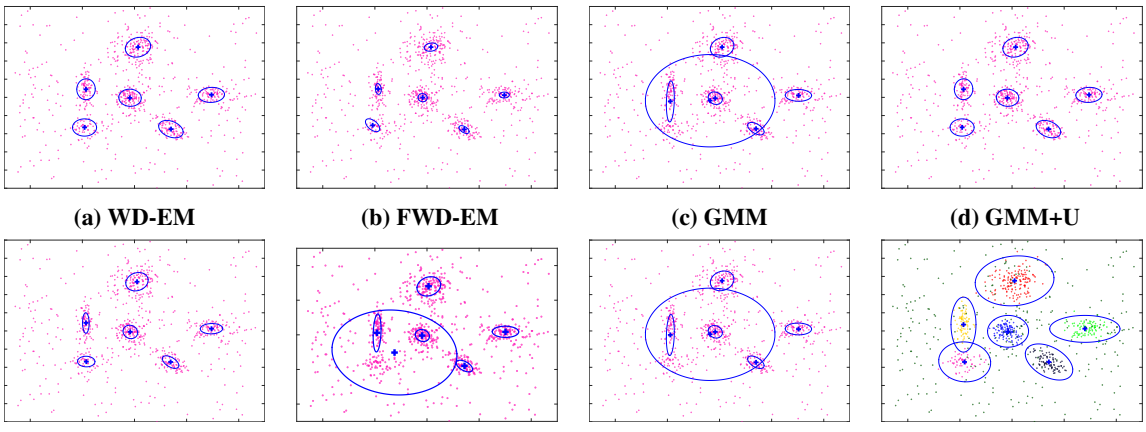
Figure 3.1: Samples of the SIM dataset with no outliers (top row) and contaminated with 50% outliers (bottom row). The 600 inliers are generated from Gaussian mixtures while the 300 outliers are generated from a uniform distribution.

Table 3.1: Datasets used for benchmarking and their characteristics: number of data points (n), dimension of the data space (d), and number of clusters (K).

Data Set	n	d	K
SIM-Easy	600	2	5
SIM-Unbalanced	600	2	4
SIM-Overlapped	600	2	4
SIM-Mixed	600	2	6
MNIST [LeCun 98]	$10,000$	141	10
Wav [Breiman 84]	$5,000$	21	3
BCW [Street 93]	569	30	2
Letter Recognition [Frey 91]	$20,000$	16	26

- **K-Means** is the standard K-means algorithm;
- **KK-Means** is the kernel K-means algorithm of [Dhillon 04];
- **NCUT** is the spectral clustering algorithm of [Shi 00].
- **HAC** is the hierarchical agglomerative clustering algorithm of [Zhao 02].

All the above algorithms need proper initialization. All the mixture-based algorithms, WD-EM, FWD-EM, GMM, GMM+U, FM-uMST, IGMM and I^{2}GMM start from the same proportions, means, and covariances which are estimated from the set of clusters provided by K-means. The latter is randomly initialized several times to find a good initialization. Furthermore, algorithms WD-EM, FWD-EM, GMM, GMM+U and FM-uMST are iterated until convergence, i.e, the log-likelihood difference between two consecutive iterations is less than 1%, or are stopped after 400 iterations.



(a) WD-EM **(b) FWD-EM** **(c) GMM** **(d) GMM+U**

(e) FM-uMST **(f) IGMM** **(g) I^{2}GMM** **(h) Ground Truth**

Figure 3.2: Results obtained by fitting mixture models to the SIM-Mixed data in the presence of 50% outliers (seeTable 3.3).

Table 3.2: Results obtained with the MNIST, WAV, BCW, and Letter Recognition datasets. The clustering scores correspond to the Davies-Bouldin (DB) index. The best results are shown in underlined **bold**, and the second best results are shown in **bold**. The proposed method yields the best results for the WAV and BCW datasets, while I^{2}GMM yields the best results for the MNIST dataset. Interestingly, the non-parametric methods (K-means, HAC and Ncut) yield excellent results for Letter Recognition.

Dataset	WD-EM	FWD-EM	GMM	GMM+U	FM-uMST	IGMM	I^{2}GMM	K-Means	KK-Means	Ncut	HAC
MNIST	2.965(0.15)	3.104(0.21)	3.291(0.14)	3.245(0.09)	**2.443(0.00)**	3.555(0.06)	**2.430(0.14)**	2.986(0.01)	2.980(0.02)	4.760(0.08)	3.178(0.00)
WAV	**0.975(0.00)**	1.019(0.00)	1.448(0.03)	1.026(0.04)	1.094(0.10)	1.028(0.02)	2.537(0.35)	1.020(0.00)	**0.975(0.05)**	2.781(0.06)	1.089(0.00)
BCW	**0.622(0.00)**	0.687(0.00)	0.714(0.00)	0.689(0.00)	0.727(0.00)	0.719(0.00)	0.736(0.09)	0.659(0.00)	**0.655(0.00)**	0.838(0.00)	0.685(0.00)
Letter Recognition	1.690(0.00)	1.767(0.01)	2.064(0.06)	2.064(0.06)	1.837(0.00)	2.341(0.11)	1.724(0.03)	**1.450(0.02)**	1.504(0.03)	**1.626(0.00)**	**1.626(0.00)**

Table 3.3: DB scores obtained on the SIM-X dataset (**best** and **second best**).

	Outliers	WD-EM	FWD-EM	GMM	GMM+U	FM-uMST	IGMM	I^{2}GMM	K-Means	KK-Means	Ncut	HAC
SIM-Easy	10%	**0.229(0.01)**	0.295(0.01)	0.295(0.01)	**0.222(0.02)**	0.307(0.02)	1.974(0.12)	0.500(0.16)	0.291(0.01)	0.330(0.07)	0.283(0.01)	0.266(0.00)
	20%	**0.266(0.02)**	0.338(0.01)	0.342(0.01)	**0.233(0.01)**	0.349(0.02)	1.564(0.43)	0.626(0.28)	0.344(0.01)	0.420(0.10)	0.335(0.01)	0.330(0.01)
	30%	**0.330(0.01)**	0.385(0.01)	0.384(0.02)	**0.227(0.02)**	0.501(0.04)	1.296(0.12)	0.570(0.27)	0.372(0.01)	0.381(0.03)	0.366(0.02)	0.376(0.01)
	40%	**0.358(0.01)**	0.445(0.04)	0.453(0.05)	**0.211(0.02)**	0.585(0.06)	1.259(0.16)	0.534(0.21)	0.417(0.01)	0.411(0.01)	0.409(0.01)	0.401(0.01)
	50%	**0.380(0.01)**	0.455(0.02)	0.459(0.02)	**0.195(0.01)**	0.568(0.05)	1.107(0.06)	0.626(0.21)	0.422(0.01)	0.439(0.03)	0.422(0.01)	0.438(0.01)
SIM-Unbalanced	10%	**0.270(0.01)**	0.954(0.72)	1.354(1.02)	**0.277(0.01)**	1.104(0.76)	1.844(0.29)	0.491(0.17)	0.405(0.02)	0.433(0.05)	0.402(0.02)	0.427(0.02)
	20%	**0.329(0.03)**	4.503(4.33)	3.003(1.85)	**0.269(0.01)**	1.181(0.44)	1.278(0.45)	0.591(0.13)	0.512(0.02)	0.515(0.03)	0.477(0.03)	0.529(0.02)
	30%	**0.399(0.03)**	3.502(3.09)	2.034(1.22)	**0.252(0.03)**	1.414(0.88)	1.272(0.35)	0.601(0.10)	0.548(0.03)	0.540(0.03)	0.531(0.02)	0.570(0.03)
	40%	**0.534(0.13)**	2.756(2.33)	2.097(1.15)	**0.251(0.02)**	1.650(0.94)	1.239(0.36)	0.615(0.05)	0.557(0.03)	0.567(0.02)	0.563(0.02)	0.597(0.02)
	50%	**0.557(0.10)**	2.400(1.44)	1.520(0.38)	**0.268(0.01)**	1.612(0.69)	1.144(0.36)	0.665(0.10)	0.580(0.03)	0.585(0.03)	0.583(0.03)	0.636(0.02)
SIM-Overlapped	10%	**0.305(0.02)**	0.693(0.31)	1.510(0.97)	**0.307(0.02)**	1.373(0.63)	2.168(0.20)	0.554(0.14)	0.395(0.03)	0.428(0.06)	0.385(0.01)	0.427(0.01)
	20%	**0.368(0.03)**	1.562(0.45)	1.881(0.50)	**0.293(0.01)**	2.702(1.28)	1.837(0.37)	0.608(0.08)	0.467(0.02)	0.532(0.07)	0.440(0.02)	0.502(0.01)
	30%	**0.472(0.04)**	1.825(0.55)	2.209(0.64)	**0.294(0.03)**	5.101(1.99)	1.568(0.61)	0.586(0.15)	0.532(0.02)	0.521(0.03)	0.508(0.01)	0.557(0.01)
	40%	0.549(0.04)	2.372(0.54)	2.597(0.73)	**0.322(0.01)**	4.569(1.72)	1.320(0.40)	0.687(0.11)	0.546(0.02)	0.556(0.03)	**0.541(0.03)**	0.593(0.02)
	50%	0.641(0.06)	2.269(0.44)	2.247(0.60)	**0.298(0.02)**	5.762(3.34)	1.174(0.25)	0.815(0.12)	0.563(0.03)	0.576(0.02)	**0.560(0.03)**	0.618(0.02)
SIM-Mixed	10%	**0.282(0.01)**	0.443(0.11)	0.448(0.11)	0.290(0.01)	0.951(0.35)	2.032(0.46)	0.414(0.12)	0.358(0.01)	0.418(0.06)	0.359(0.01)	0.355(0.01)
	20%	**0.351(0.02)**	0.857(0.52)	1.325(0.79)	**0.286(0.01)**	1.062(0.38)	1.782(0.44)	0.462(0.08)	0.413(0.02)	0.476(0.06)	0.409(0.01)	0.428(0.01)
	30%	**0.396(0.02)**	1.368(0.74)	1.524(0.64)	**0.278(0.01)**	1.693(0.56)	1.627(0.54)	0.483(0.07)	0.454(0.02)	0.464(0.04)	0.449(0.01)	0.468(0.01)
	40%	**0.449(0.03)**	1.100(0.61)	1.188(0.59)	**0.277(0.02)**	1.609(0.43)	1.456(0.34)	0.483(0.05)	0.478(0.02)	0.504(0.04)	0.478(0.01)	0.508(0.02)
	50%	**0.492(0.03)**	1.364(0.59)	1.513(0.67)	**0.265(0.01)**	1.972(0.86)	1.366(0.29)	0.562(0.04)	0.501(0.01)	0.515(0.02)	0.499(0.02)	0.546(0.02)

To quantitatively evaluate all the tested methods, we chose to use the Davies-Bouldin (DB) index [Davies 79]:

$$DB = \frac{1}{K}\sum_{k=1}^{K} R_k, \tag{3.40}$$

where $R_k = \max_{k,k\neq l}\{(S_k + S_l)/d_{kl}\}$, $S_k = n_k^{-1}\sum_{\boldsymbol{x}\in C_k}\|\boldsymbol{x} - \boldsymbol{\mu}_k\|$ is the cluster scatter, n_k is the number of samples in cluster k, $\boldsymbol{\mu}_k$ is the cluster center, and $d_{kl} = \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|$. A low value of the DB index means that the clusters are far from each other with respect to their scatter, and therefore the discriminative power is higher. Since the algorithms are randomly initialized, we repeat each experiment 20 times and compute the mean and standard deviation of the DB index for each experiment. Table 3.2 summarizes the results obtained with the MNIST, WAV, BCW, and Letter Recognition datasets. The proposed WD-EM method yields the best results for the WAV and BCW data, while the I^{2}GMM yields the best result for the MNIST data. It is interesting to notice that the non-parametric methods K-means, NCUT and HAC yield the best and second best results for the Letter Recognition data.

An interesting feature of the proposed weighted-data clustering algorithms is their robustness in finding good clusters in the presence of outliers. To illustrate this ability we run a large number of experiments by adding outliers, drawn from a uniform distribution, to the four simulated datasets, e.g., Table 3.3 and Fig. 3.2. A comparison between WD-EM, FWD-EM, and the state-of-art clustering techniques mentioned above, with different percentages of outliers, is provided. As it can be easily observed in these tables, GMM+U performs extremely well in the presence of outliers, which is not surprising since the simulated outliers are drawn from a uniform distribution. Overall, the proposed WD-EM method is the second best performing method. Notice the very good performance of the Ncut method for the SIM-overlapped data. Among all these methods, only GMM+U and WD-EM offer the possibility to characterize the outliers using two very different strategies.The GMM+U model simply pulls them in an *outlier class* based on the posterior probabilities. The WD-EM algorithm iteratively updates the posterior probabilities of the weights, and the final posteriors, (3.19), allow to implement a simple outlier detection mechanism. Another important remark is that WD-EM systematically outperforms FWD-EM, which fully justifies the proposed weighted-data model. Fig. 3.2 shows results of fitting the mixture models to SIM-mixed data drawn from a Gaussian mixture and contaminated with 50% outliers drawn from a uniform distribution. These plots show that GMM, IGMM, and I^{2}GMM find five components corresponding to data clusters while they also fit a component onto the outliers, roughly centered on the data set.

3.9 CONCLUSIONS

We presented a weighted-data Gaussian mixture model. We derived a maximum-likelihood formulation and we devised two EM algorithms, one that used fixed weights (FWD-EM) and another one with weights modeled as random variables (WD-EM). While the first algorithm appears to be a straightforward generalization of standard EM for Gaussian mixtures, the second one has a more complex structure. We showed that the expectation and maximization steps of the proposed WD-EM admit closed-form solutions and hence the algorithm is extremely efficient. Moreover, WD-EM performs much better than FWD-EM which fully justifies the proposed generative probabilistic model for the weights. We extended the MML-based model selection criterion proposed in [Figueiredo 02] to the weighted-data Gaussian mixture model and we proposed an algorithm that finds an op-

timal number of components in the data. Interestingly, the WD-EM algorithm compares favorably with several state-of-the-art parametric and non-parametric clustering methods: it performs particularly well in the presence of a large number of outliers, e.g., up to 50% of the inlier data. Hence, it can be referred to as robust clustering.

Algorithm 1: WD-EM with model selection based on the MML criterion.

input : $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n, K_{\text{low}}, K_{\text{high}}, \boldsymbol{\Theta}^{(0)} = \{\pi_k^{(0)}, \boldsymbol{\mu}_k^{(0)}, \boldsymbol{\Sigma}_k^{(0)}\}_{K=1}^{K_{\text{high}}}, \boldsymbol{\Phi}^{(0)} = \{\alpha_i^{(0)}, \beta_i^{(0)}\}_{i=1}^n$

output: The minimum length mixture model: $\boldsymbol{\Theta}_{\min}$ and the final data weights: $\boldsymbol{W}_{\min}$

Set: $r = 0, \mathcal{K}^+ = \{k\}_{k=1}^{K_{\text{high}}}, \mathbf{LEN}_{\min} = +\infty$

while $|\mathcal{K}^+| \geq K_{\text{low}}$ **do**

  repeat

    for $k = 1$ **to** K_{high} **do**

      E-Z step using (3.17):

$$\eta_{ik}^{(r+1)} = \frac{\pi_k^{(r)} \mathcal{P}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_k^{(r)}, \boldsymbol{\Sigma}_k^{(r)}, \alpha_i^{(r)}, \beta_{ik}^{(r)}\right)}{\sum_{l=1}^{K_{\text{high}}} \pi_l^{(r)} \mathcal{P}\left(\boldsymbol{x}_i; \boldsymbol{\mu}_l^{(r)}, \boldsymbol{\Sigma}_l^{(r)}, \alpha_i^{(r)}, \beta_{il}^{(r)}\right)}$$

      E-W step using (3.20)–(3.21):

$$\alpha_i^{(r+1)} = \alpha_i^{(0)} + \frac{d}{2}$$

$$\beta_{ik}^{(r+1)} = \beta_i^{(0)} + \frac{1}{2} \left\| \boldsymbol{x}_i - \boldsymbol{\mu}_k^{(r)} \right\|_{\boldsymbol{\Sigma}_k^{(r)}}^2$$

$$\overline{w}_{ik} = \frac{\alpha_i^{(r+1)}}{\beta_{ik}^{(r+1)}}$$

      M-step

$$\pi_k^{(r+1)} = \frac{\max\left\{0, \sum_{i=1}^n \eta_{ik}^{(r+1)} - \frac{M}{2}\right\}}{\sum_{l=1}^{K_{\text{high}}} \max\left\{0, \sum_{i=1}^n \eta_{il}^{(r+1)} - \frac{M}{2}\right\}}$$

      if $\pi_k^{(r+1)} > 0$ **then**

        Evaluate $\boldsymbol{\theta}_k^{(r+1)}$: mean $\boldsymbol{\mu}_k^{(r+1)}$ using (3.27) and covariance $\boldsymbol{\Sigma}_k^{(r+1)}$ using (3.28).

      else

        $K^+ = K^+ - 1$

      end

    end

    $\boldsymbol{\Theta}^{(r+1)} = \left\{\pi_k^{(r+1)}, \boldsymbol{\theta}_k^{(r+1)}\right\}_{k=1}^{K_{\text{high}}}$

    Compute optimal length $\mathbf{LEN}_{\text{MML}}^{(r+1)}$ with (3.37).

    $r \leftarrow r + 1$

  until $\left|\Delta \mathbf{LEN}_{MML}^{(r)}\right| < \varepsilon$

  if $\mathbf{LEN}_{MML}^{(r)} < \mathbf{LEN}_{\min}$ **then**

    $\mathbf{LEN}_{\min} = \mathbf{LEN}_{\text{MML}}^{(r)}$

    $\boldsymbol{\Theta}_{\min} = \boldsymbol{\Theta}^{(r)}$

    $\boldsymbol{W}_{\min} = \{\overline{w}_i\}_{i=1}^n$ with $\overline{w}_i = \sum_{k=1}^{K_{\text{high}}} \eta_{ik} \overline{w}_{ik}$

  end

  $k^* = \text{argmin}_{k' \in \mathcal{K}^+} \left(\pi_{k'}^{(r)}\right), \quad \mathcal{K}^+ = \mathcal{K}^+ / k^*$

end

CHAPTER 4

WEIGHTED DATA CLUSTERING APPLIED TO AV SPEAKER LOCALIZATION

In this chapter we address the task of audio-visual speaker localization, *i.e.,* given audio and video streams of a scene (say from multiple microphones and cameras), one would like to know if someone is speaking and where that person is. A robust solution to this problem is useful for many applications that require the location of speakers as an input, *e.g.,*, audio-visual tracking, speaker diarization, dialogue modeling, etc. We propose to address this in the framework of data clustering. First, speaker related features are extracted from the audio and video stream gathered with a single camera and two microphones. Then the weighted-data clustering algorithm (WD-EM) proposed in Chapter 3 is applied to meaningfully and efficiently fuse and cluster features from the two modalities. We performed experiment on a publicly available dataset. The results we obtained show that multimodal data processing compensates for the weaknesses of visual-only or audio-only data analysis.

4.1 INTRODUCTION

In recent years, audio-visual scene analysis has drawn increasing attention of researchers working in the field of signal processing and machine learning, primary due to its potential to provide rich information that can be useful in a number of applications *e.g.,* robotics, surveillance, social computing, etc. In typical real-world untethered human-robot interaction scenarios, humans are at some distance from the robot and the acquired signals, e.g., by one or more microphone and video camera embedded in the robot platform, are processed to enable the robot to navigate in complex environments and act as social, cognitive human partner. However, enabling robots with multimodal perception of their environment is an essential and challenging task.

In this chapter, amongst the possible audio-visual analysis tasks, we focus on the task of locating speakers in informal scenarios. This problem arises when the task is, *e.g.,*

to detect a person that is both seen and heard, such as an active speaker in the field of view (FoV) of a humanoid robot. In other words we are interested in retrieving the active speakers in a group of people engaged in a natural social interplay, by means of auditory and visual information. Example of such scenarios in which two or more people are sitting and chatting in front of the robot are shown in Fig. 4.1.

Recently, multimodal speaker localization techniques that use audio and video information have been shown to effectively address some of the problems in audio-only and video-only systems. But the challenge is how to effectively perform fusion so that one modality may compensate for the weakness of the other one. Single-modality signals vision-only or audio-only are often either weak or ambiguous, and it may be useful to combine information from different sensors, *e.g.,* cameras and microphones. However, there are several difficulties associated with audio-visual fusion, i.e., the two sensorial modalities: (i) live in different mathematical spaces or have a different representation, (ii) are contaminated by different types of noise with different distributions, (iii) are perturbed by different physical phenomena, *e.g.,* acoustic reverberations, ambient noise, lighting conditions, etc, and (iv) have different spatio-temporal resolution.

Moreover, a speaker may face the camera while he/she is silent and may emit speech while he/she turns his/her face away from the camera. Speech signals have sparse spectral and temporal structure and they are mixed with other sound sources, such as music or background noise. In visual modality, speaker faces may be totally or partially occluded, in which case face detection and localization is extremely unreliable. We note that it is necessary to fuse the information from both modalities in a way that the complementary information is exploited to solve the problem at hand.

In this chapter, we present a robust and instantaneous active speaker(s) localization framework based the fusion of audio and video data. We present a methodology in which pieces of information extracted from audio and vision modalities are weighted accordingly to their relevance for speaker localization task. We show that the proposed method is well suited to find audio-visual clusters and to discriminate between speaking and silent people. One limitation of our proposed framework is that it requires a unique hardware setup: a video camera and a binaural microphone are needed.

The reminder of the chapter is structured as follows. Section 4.2 delineates the related published work. Section 4.3 describes the details of the auditory and visual extracted features and data clustering framework. In Section 4.4 we shows experimental result obtained on real audio-visual recording. Section 4.5 ends the chapter by presenting some conclusions and suggestion for further work.

4.2 Related Work

Among the different methods that perform speaker localization, only a few are performing the fusion of both audio and video modalities. Earlier research were based on finding synchrony between audio and video streams. They were inspired from the observation that video and audio events tends to occur together; not always but often; lips moving and speech when talking; the movements of fingers and the sound of instrument when

Figure 4.1: A typical scenario in untethered human-robot-interaction. A companion humanoid robot (NAO) perform audio-visual scene analysis in an attempt to detect and locate speakers in the field of view of its camera.

a piano, guitar or drum is played. The audio and video events are concurrent in these cases because there is a common physical cause. The authors in [Hershey 00] presents a method to locate sound source in the image, based on quantifying the synchrony between the auditory and the visual modalities. The most audio synchronized region of the video frame was selected as source of the corresponding audio clip (speaker).

The work in [Hershey 00] inspired a series of information-theory based papers. For example, a statistical framework to measure the amount of mutual information between a region of interest on the image and the audio track is proposed in [Fisher 04]. The works in [Butz 02, Beal 02, Besson 08] follow a similar approach to determine the active speaker among a few candidate faces. The main advantage of these approaches is the versatility, since they are not constrained to a particular kind of objects. However, they require high-resolution images acquired with speaker-dedicated cameras, i.e., frontal facing speaker to the video camera. Therefore, their use is restricted mostly to static scenarios when the number of speakers is constant and known in advance and the background scene is static with no distractive motion. Thus, they can no be applied in a general audio-visual scene.

In [Talantzis 08, Zotkin 02, Checka 04] Bayesian frameworks inferring the position of the active speaker by combining a sound source localization technique with a face tracking algorithm were proposed. However, any tracking system has two main potential problems: initial position of the tracks and tracker drift. Indeed, often, tracking methods need an accurate guess of the track start as well as a robust instantaneous detector avoiding the tracker to drift apart from the true position. Therefore, the development of a robust and instantaneous speaker localization method is the best way to enhance the performance of the current tracking methodologies. In [Noulas 12] speaker diarization model that capture the casual relationship between the speaker and the multi-modal recording she/he produces is proposed. A set of observations from audio and video modality and their joint space (e.g, the motion estimation of face region and the energy variation of audio over time) are used to infer the image location of the speaker on each video frame. The authors demonstrated the model on a dataset of meeting recordings with front facing camera available for each participant.

Several probabilistic approaches dealing with the instantaneous localization of speakers have been published [Khalidov 11b, Khalidov 08b, Alameda-Pineda 11]. The common point of these studies is that they cast the speaker localization problem into a mul-

timodal clustering task. In that sense, our work in this chapter is inspired from them. More precisely, [Khalidov 11b, Khalidov 08b] use two Gaussian Mixture Models (GMM) one per modality. The parameters of the two GMM are constrained via a subset of tying parameters. The resulting EM algorithm has a computationally expensive M step, involving non-linear optimization subroutines, due to the parameters' constraints. In [Alameda-Pineda 11] a single GMM is used to cluster multimodal data and the mixture parameters are estimated relaying more on visual than on auditory data.

None of the methods above addresses the problem of audio-visual fusion with weighted data. Indeed, most of them are able to trust one of the two modalities, but none is able to give a different weight to the observations coming from one modality. Even if weighted-data clustering has been addressed in the recent past [Xi 04, Forbes 10], up to our knowledge this work is the very first study on how to use weights on multimodal data clustering. The main contribution of this chapter is a robust and instantaneous method for active speakers localization that combine information from audio and visual modalities. More importantly, we present a methodology with the following remarkable attributes all together: (i) a methodology in which the pieces of information are weighted accordingly to their relevance to solve the task at hand, (ii) unlike other audio-visual fusion method that uses spatio-temporal visual features, we only use visual information that are related to audio, i.e, face and lips location information, (iii) can handling a variable number of people in unrestricted indoor environments.

Our most revealing contribution to the field is that the weighted-data clustering method proposed in chapter 3 is well suited to find audio-visual clusters and to discriminate between speaking and silent people. We showed that the proper use of these weights can notably increase the performance of speaker localization task.

4.3 AUDIO-VISUAL CLUSTERING

We cast the active speaker localization problem into a data clustering task and develop an instantaneous speaker localization framework that uses both auditory and visual information. However, prior to clustering one needs to represent data from auditory and visual modalities in the same euclidean space. Section 4.3.1 describes the steps auditory observation are extracted from the binaural audio stream and section 4.3.2 describes the steps for extracting visual observations from the video stream. To this end, let's assume we have the auditory and visual data in the same euclidean space. We then apply the **WD-EM** algorithm proposed in chapter 3 to fit a GMM model. Cross-modal weights are used to systematically provide a relevance measure for each modality observation in a data driven fashion. That is to say, the auditory observation are used to weight the visual observations and vise-versa, as explained in section 4.3.3. Model selection procedure proposed in section 3.6 is applied to find the number of gaussian component that best fit the set of multimodal observations, this also corresponds to finding the number of speakers.

4.3.1 THE AUDIO MODALITY

Extracting meaningful features from the auditory signals acquired at one or more microphones is a difficult task for several reasons. First, the different auditory channels

are contaminated by noise coming from the microphones and reverberations, which can highly perturb the signal. Second, the information we need for our task, e.g., the position of speaker or the sound source, is embedded in the different auditory channels in a complex and environment-dependent fashion. Third, the information is sparsely distributed in the auditory signal, both in time and frequency and it is only meaningful when the sound sources are active.

In order to provide to the proposed model reliable auditory features, without loss of generality we adopt the sound-source localization method of [Deleforge 13] that performs direction of arrival (DOA) estimation in 2-D followed by mapping the estimated sound-source direction onto the image plane: a DOA estimate therefore corresponds to a pixel location in the image plane. We chose the sound source localization method proposed in [Deleforge 13] for its performance and robustness. The method uses spectral binaural cues to learn the effect of the environment on the acoustic signals and therefore it is able to accurately find the sound source position. More precisely, the method requires a training phase with white noise in which the position of the sound source is known during the extraction of the spectral binaural cues (Interaural Phase and Level Differences). These position-cue pairs are used to learn a probabilistic mapping from the source position space to the spectral domain. Moreover, the probabilistic framework provides the inversion mapping, thus a sound source localization mapping from spectral binaural cues. A prominent feature of the probabilistic model is the explicit modeling of *missing data* situations. That is to say, that the localization mapping does not need a spectral cue with meaningful information in all frequency bands. Instead, the mapping makes use of those frequency bands in which the source is emitting, and is able to decouple the content of the sound source from its position. Therefore, we find that the method proposed in [Deleforge 13] is extremely well adapted to the scenario of our research.

In practice, we train the method with a loudspeaker emitting white noise and carrying an easy-to-detect visual target. This target provides the image location associated to the extracted binaural cues. Once the localization mapping is trained, during test time, given audio stream we use it to extract potential source locations that will be denoted by $\boldsymbol{A} = \{\boldsymbol{a}_j\}_{j=1}^{n_a}$, here after referred as auditory observations.

4.3.2 The Visual Modality

Together with the signals acquired at the microphones, we also use the image flows captured by a color camera. Visual information is less sparse than auditory information. As long as the speakers are in the field of view and they are not occluded, it is possible to get video location of possible speakers in an enclosed scenario.

As in the previous section, we would like to provide to the proposed model features that robustly localize the people in the visual field of view. One may immediately think about detection the speakers' face. However, this has shown to be a limitation in many previous works [Besson 08, Fisher 04, Noulas 12], in which non-frontal detection was not possible. Instead, we first detect human upper body using [Ferrari 08]. This detector provides an approximate location of the head. In order to refine this localization, we run

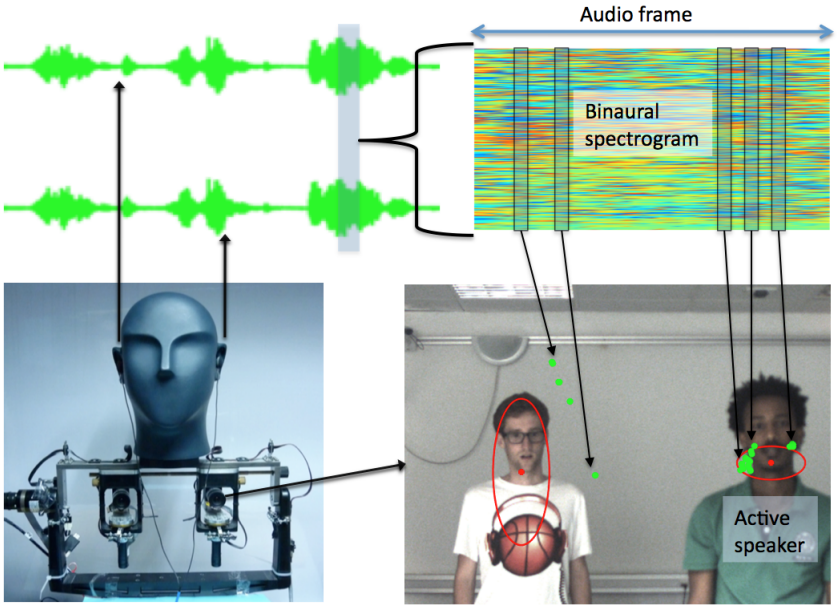


Figure 4.2: The auditory and visual data are recorded with two microphones and one camera. The audio signals are segmented into frames and each frame (vertical grey rectangle) is transformed into a binaural spectrogram. This spectrogram is composed of a sequence of binaural vectors (vertical rectangles) and each binaural vector is mapped onto a sound-source direction which corresponds to a point in the image plane (green dots).

the landmark face detector presented [Zhu 12]. One of the prominent features of this method is that it provides position of the lip landmarks. Therefore, if a face is found, the head position is replaced by the average position of the lip landmarks. In this way we build a general-purpose visual person localizer that is robust to light changes and to pose thanks to [Ferrari 08, Zhu 12] and that always provides a localization, refined in the case of frontal detection. From now on, these localizations will be denoted by $\boldsymbol{V} = \{\boldsymbol{v}_l\}_{l=1}^{n_v}$, here after referred as visual observations.

4.3.3 CROSS-MODAL WEIGHTING

As discussed in Section 3.7, we need to provide the prior parameters Φ of the weights $\boldsymbol{W}$ associated to the audio-visual observations $\boldsymbol{X} = \boldsymbol{A} \cup \boldsymbol{V}$. From now on we write $\boldsymbol{x}_i = \boldsymbol{a}_i$ for $i = 1, \ldots, n_a$ and $\boldsymbol{x}_i = \boldsymbol{v}_{i-n_a}$ for $i = n_a + 1, \ldots, n = n_a + n_v$. In other words, the first n_a are auditory observations and the remaining n_v are the visual observations.

In this context the following natural question arises: how can we systematically provide values for Φ in a data-driven fashion and that will help the EM algorithm group the observations? Intuitively, we would like auditory observations that are close to visual observations to have higher relevance that those auditory observation lying far away from all visual observations. The same intuition hold for visual observations that are close/far from auditory observations. The rationale behind this choice is that one auditory observation far away from all visual observations is probably an outlier. However, when an auditory observation is close to many visual observations, there is a bigger chance that it

corresponds to an underlying audio-visual cluster (a speaker). Therefore, the latter kind observations should have larger weight than the former kind of observations. In order to make this intuition real we compute the following quantity for each observation $\boldsymbol{x}_i$:

$$w_i^{(0)} = \sum_{s \in \mathcal{S}_i} \exp\left(-\frac{D^2(\boldsymbol{x}_i, \boldsymbol{x}_s)}{\sigma}\right),$$

where D is a distance function, e.g., euclidean distance. In the previous formula, $\mathcal{S}_i = \{1, \ldots, n_a\}$ if $i > n_a$ and $\mathcal{S} = \{n_a + 1, \ldots, n_v\}$ if $i \leq n_a$. That is to say that we use the visual observations to compute the weight for the $\boldsymbol{a}_j$'s and the auditory observations to compute the weight for the $\boldsymbol{v}_l$'s. The parameters of the prior gamma distribution are set to $\alpha_i = w_i^{(0)2}$ and $\beta_i = w_i^{(0)}$. In this way, the mode and variance of the prior distribution for w_i are $w_i^{(0)}$ and 1 respectively. One can notice, this *cross-modal* weighting scheme favors clusters composed of both auditory and visual observations.

4.3.4 Determining the Number of Speakers

In our particular application, we do not know the number of speakers beforehand. In order to overcome this issue, we used WD-EM algorithm with the Minimum Message Length (MML) criterion as presented in section 3.6 for model selection. The cross-modal weighting together with the MML criterion creates a robust method to coherently group auditory and visual observations.

4.3.5 Post Processing

We assume that each cluster found when the EM converge represents a potential speaker. However, the present clustering framework is application-blind. The model best fitting the auditory and visual observations does not necessarily correspond to the best representation of the ongoing social interplay. In our particular case, this translates into getting spurious groups of observations (clusters) that do not correspond to a speaker in the scene. More precisely, we may have three type of clusters: (i) groups of auditory observations that do not contain any visual observations, (ii) groups of only visual observations, and (iii) groups containing both audio and visual observations. In the first case, the cluster should be discarded, since the probability of a systematic fail of the upper-body detector is very low. In particular, this cluster may represent non-speech audio source which we are not interested in this work or a group of outlier audio points. In the second case, we could keep the cluster and mark it as a potentially silent speaker.

We are mostly interested in third type of clusters that contain both auditory and visual observations. With this aim, we classify all the observations into clusters using MAP. Clusters containing both video observations and a sufficient number of audio observations are marked as active speakers. By sufficient we mean no less than $\frac{n_a+n_v}{\hat{K}}$, where $\hat{K}$ is the number of cluster chosen by MML. We found this value high enough to discard clusters containing auditory outliers and small enough to guarantee the good sensibility of the system.

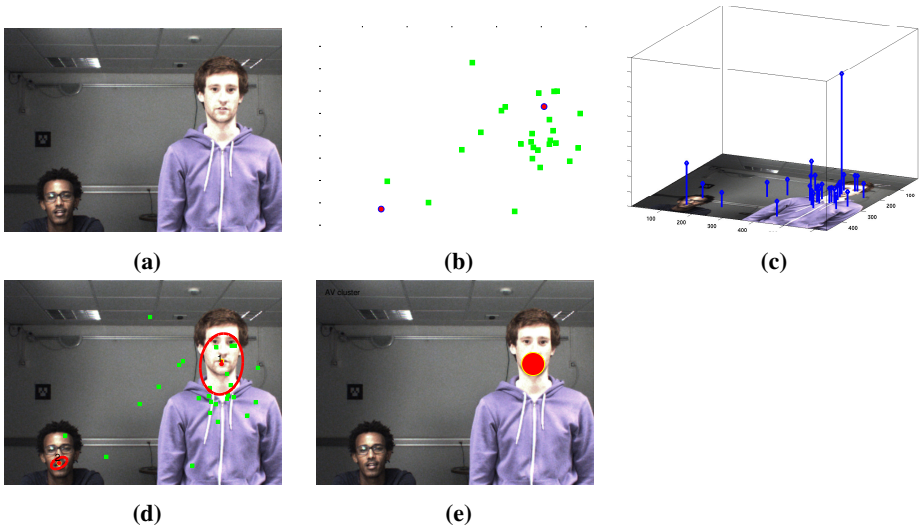


Figure 4.3: Sample frame from *fake speaker* (FS) sequence to demonstrate the application of WD-EM for speaker localization. (a) a sample video frame, in this case the right person is speaking, (b) audio observation (in green dots) and visual observation (in red dots), (c) initial weight of observation as a stem plot, (d) two clusters are found by model selection, (e) the active speaker is found after post-processing procedure is shown with a red filled circle

4.4 EXPERIMENTS

4.4.1 DATA COLLECTION

To illustrate the effectiveness of both the data clustering model proposed in chapter 3 and the speaker localization framework outlined in the previous section, we recorded three sequences:

- The *fake speaker* (FS) sequence, *e.g.,* first and second rows of Fig. 4.4, consists of two persons facing the camera and the microphones. While the person onto the right emits speech signals (counting from “one" to “ten") the person onto the left performs fake lip, facial, and head movements as he would speak.
- The *moving speakers* (MS) sequence, *e.g.,* third and fourth rows of Fig. 4.4, consists of two persons that move around while they are always facing the cameras and microphones. The persons take speech turns but there is a short overlap between the two auditory signals.
- The *cocktail party* (CP) sequence, *e.g.,* fifth and sixth rows of Fig. 4.4, consists of four persons engaged in an informal dialog. The persons wander around and turn their heads towards the active speaker; occasionally two persons speak simultaneously. Moreover the speakers do not always face the camera, hence face and lip detection/localization are unreliable.

The visual data are gathered with a single camera and the auditory data are gathered with two microphones plugged into the ears of an acoustic dummy head, referred to as *binaural*

audition. The visual data are recorded at 25 video frames per second (FPS). The auditory data are gathered and processed in the following way. First, the short-time Fourier transform (STFT) is applied to the left- and right-microphone signals which are sampled at 48 KHz. Second, the left and right spectrograms thus obtained are combined to yield a binaural spectrogram from which a sound-source DOA is estimated. A spectrogram composed of 512 frequency bins is obtained by applying the STFT over a sliding window of width 0.064 s and shifted along the signal with 0.008 s hops. An audio frame, or 512 frequency bins, is associated with each window, hence there are 125 audio frames per second (with 0.056 ms overlap between consecutive frames). Both the visual and audio frames are further grouped into temporal segments of width 0.4 s, hence there are 10 visual frames and 50 audio frames in each segment.

As already mentioned, we follow the method of [Deleforge 15a] to extract a sound-source DOA from each audio frame. In order to increase the robustness of audio localization, a voice activity detector (VAD) [Sohn 99] is first applied to each frame, such that not all the frames have DOA estimates associated with them. On an average there are 40 audio DOA observations per segment. The FS sequence contains 28 segments, the MS sequence contains 43 segments, while the CP sequence contains 115 segments. The left hand sides of Fig. 4.4 show the central frame of a segment with all the visual features (blue) and auditory features (green) available within that segment.

4.4.2 RESULTS

We tested the proposed WD-EM algorithm on these audio-visual data as well as the GMM+U [Banfield 93] and FM-uMST [Lee 14] algorithms. We chose to compare our method with these two methods for the following reasons. Firstly, all three methods are based on finite mixtures and hence they can use a model selection criterion to estimate the number of components in the mixture that best approximates clusters in the data. This is important since the number of persons and of active speakers among these persons are not known in advance. Secondly, as demonstrated in the previous section, these three methods yield robust clustering in the presence of outliers.

WD-EM uses the MML criterion for model selection as described in Section 3.6. We implemented a model selection criterion based on BIC to optimally select the number of components with GMM+U and FM-uMST. While each algorithm yields an optimal number of components for each audio-visual segment, not all them contain a sufficient number of audio and visual observations, such that the component can be associated with an active speaker. Therefore, we apply a simple two-step strategy, firstly to decide whether a component is *audio-visual*, *audio-only*, or *visual-only*, and secondly to select the best audio-visual components. Let n_v and n_a be the total number of visual and audio observations in a segment. We start by assigning each observation to a component: let n_a^k and n_v^k be the number of audio and visual observations associated with component k. Let $r_k = \min\{n_a^k, n_v^k\}/(n_a + n_v)$ measure the audio-visual relevance of a component. If $r_k \geq s$ then component k corresponds to an active speaker, with s being a fixed threshold.

Fig. 4.4 shows examples of applying the WD-EM, GMM+U and FM-uMST algorithms to the three sequences. One may notice that, while the visual observations (blue) are very

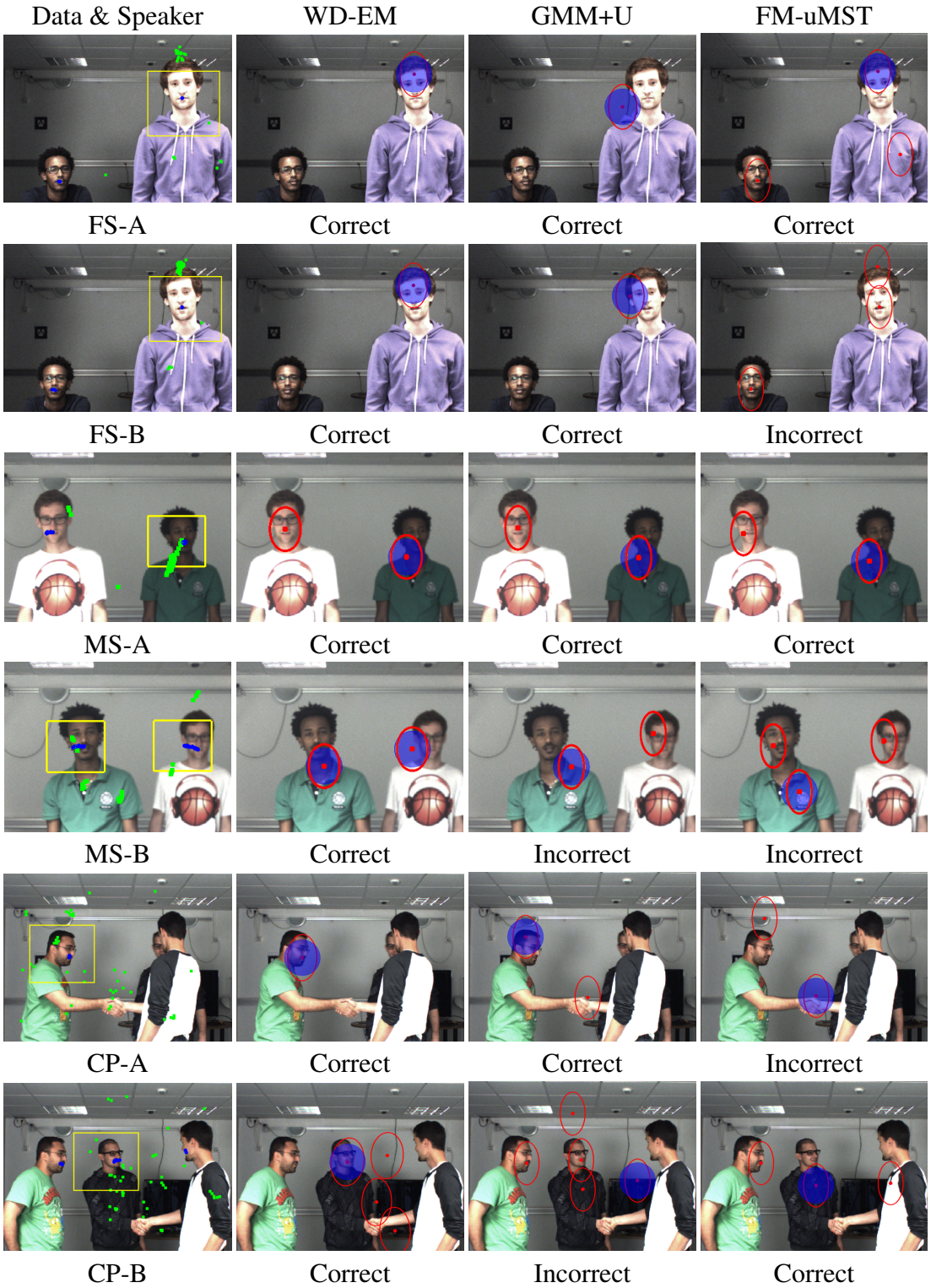


Figure 4.4: Results obtained on the *fake speaker* (FS), *moving speaker* (MS) and *cocktail party* (CP) sequences. The first column shows the audio (green) and visual (blue) observations, as well as a yellow bounding box that shows the ground-truth active speaker. The second, third and fourth columns show the mixture components obtained with the WD-EM, GMM+U and FM-uMST methods, respectively. The blue disks mark components that correspond to correct detections of active speakers, namely whenever there is an overlap between a component and the ground-truth bounding box.

Table 4.1: The correct detection rates (CDR) obtained with the three methods for three scenarios: fake speaker (FS), moving speakers (MS), and cocktail party (CP).

Scenario	# Segments	WD-EM	GMM-U [Banfield 93]	FM-uMST [Lee 14]
FS	28	100.00%	100.00%	71.43%
MS	43	83.87%	61.90%	72.22%
CP	115	65.66%	52.48%	49.57%

accurate and form small *lumps* around the moving lips of a speaker (or of a fake speaker), audio observations (green) are very noisy and have different statistics; this is due to the presence of reverberations (the ceiling in particular) and of other sound sources, such as computer fans. The ground-truth active speaker is shown with a yellow frame. The data clusters obtained by the three methods are shown with red ellipses. A blue disk around a cluster center designates an audio-visual cluster. Altogether, one may notice that the proposed method outperforms the two other methods. An interesting feature of WD-EM is that the weights give more importance to the accurate visual data (because of the low-variance groups of observations available with these data) and hence the audio-visual cluster centers are pulled towards the visual data (lip locations in these examples).

To further quantify the performance of the three methods, we carefully annotated the data. For each segment, we identified the active speaker and we precisely located the speaker's lips. Let $\boldsymbol{x}_g$ be the ground-truth lip location. We assign $\boldsymbol{x}_g$ to a component by computing the maximum responsibility(3.17) of $\boldsymbol{x}_g$. When $\boldsymbol{x}_g$ is assigned to an audio-visual cluster, an active speaker is said to be correctly detected if the posterior probability of $\boldsymbol{x}_g$ is equal or greater than $1/K$, where K is the number of components. Table 4.1 summarizes the results obtained with the three methods.

4.5 CONCLUSIONS

One of the important tasks in many audio-visual analysis application is active speaker localization. In this chapter we addressed this problem in the framework of data clustering. We briefly described the audio-visual speaker localization problem and how it may be cast into a challenging audio-visual data clustering problem, *e.g.,* how to associate human faces with speech signals and how to detect and localize active speakers in complex audio-visual scenes. We applied WD-EM algorithm to meaningful and efficiently cluster the audio-visual data. We showed that for the task at hand the proposed algorithm yields better audio-visual clustering results than two other finite-mixture models, and this for two reasons: (i) it is very robust to noise and to outliers and (ii) it allows a cross-modal weighting scheme, which is an important feature for multimodal applications. Although not implemented in this chapter, the proposed model has many other interesting features when dealing with multimodal data: it enables to balance the importance of the modalities, to emphasize one modality, or to use any prior information that might be available, for example by giving high weight priors to visual data corresponding to face/lip localization. This is left for future research.

CHAPTER 5

AV TRACKING BY DENSITY APPROXIMATION IN A SEQUENTIAL BAYESIAN FILTERING FRAMEWORK

This chapter presents a novel audio-visual tracking approach that exploits constructively audio and visual modalities in order to estimate trajectories of multiple people in a joint state space. The tracking problem is modeled using a sequential Bayesian filtering framework. Within this framework, we propose to represent the posterior density with a Gaussian Mixture Model (GMM). To ensure that a GMM representation can be retained sequentially over time, the predictive density is approximated by a GMM using the Unscented Transform. While a density interpolation technique is introduced to obtain a continuous representation of the observation likelihood, which is also a GMM. Furthermore, to prevent the number of mixtures from growing exponentially over time, a density approximation based on the Expectation Maximization (EM) algorithm is applied, resulting in a compact GMM representation of the posterior density. Recordings using a camcorder and microphone array are used to evaluate the proposed approach, demonstrating significant improvements in tracking performance of the proposed audio-visual approach compared to two benchmark visual trackers.

5.1 INTRODUCTION

Awareness of the surrounding environment is a prerequisite for interaction between humans and autonomous systems. In particular for HRI, knowledge of the directions of sound sources in the surrounding acoustic environment is crucial in order to approach, look at, focus on, and engage with users. However, in realistic conditions, speech radiated in enclosed environments is subject to reverberation due to reflections off surrounding walls and objects. Dominant early reflections therefore often lead to spurious false detections, whilst late reverberation causes localization errors. Furthermore, human talkers are highly dynamic sources, such that DOAs are both spatially- and time-varying. Therefore,

source tracking approaches constructively exploit temporal models of the source dynamics to estimate and smooth the trajectory of corresponding DOAs. Whilst acoustic source tracking approaches, such as [Evers 15], propagate source trajectories through natural periods of speech inactivity within sentences, prolonged inactivity during dialogues often leads to track deletions.

Nevertheless, especially in the case of robotics, acoustic sensors are often coupled with synchronized camera systems. Therefore, visual information can be exploited constructively to disambiguate the practical challenges of acoustic signal processing. Recent contributions in the audio-visual community therefore utilize features extracted from images to complement audio processing tasks and *vice versa*. The attention control system for mobile robots in [Lang 03, Sanchez-Riera 12] uses separate but parallel audio and visual processing subsystems to identify salient events. In [Naqvi 10b], a visual tracker is used to estimate the positions and velocities of people for blind source separation. Furthermore, a multi-person visual tracker is used for speaker diarization in multi-party dialogues in [Gebru 17a].

Nevertheless, to fully exploit the information of both modalities, audio-visual fusion – rather than disambiguation – is necessary. An audio-visual tracking system is proposed in [Kılıç 15b, Gatica-Perez 07] that estimates the trajectories of talkers from the joint signals of several distributed cameras and a large microphone array. Nevertheless, joint audio-visual tracking approaches for compact configurations, such as for robot audition, are a novel contribution to the literature. In sequential Bayesian frameworks, the primary challenge for joint estimation from both modalities is that audio-visual observations are typically highly non-linear, non-Gaussian and multi-modal. The posterior Probability Density Function (pdf) is therefore analytically intractable. Classical approximations for non-linear systems, such as the Extended Kalman Filter (EKF) [Anderson 79] and Unscented Kalman Filter (UKF) [Julier 97] are only valid for the estimation and propagation of unimodal Gaussian densities. Nevertheless, the posterior pdf for audio-visual tracking is highly multi-modal and potentially heavy tailed. Approaches using particle filters (PF) and Monte Carlo Markov chains (MCMC) were shown to be an effective alternative to Kalman filter variants in order to approximate the posterior density of non-linear and non-Gaussian dynamic systems. However, significant amount of particles are typically required for good approximation and computational cost can be prohibitive [Andrieu 03].

In this chapter, we therefore propose to approximate the posterior density by a Gaussian Mixture Model (GMM). To ensure that a GMM representation can be retained over time, the predictive density is approximated by a GMM using the Unscented Transform (UT) [Van Der Merwe 00]. The UT calculates the statistics of a random variable which undergoes a non-linear transformation and builds on the principle that it is easier to approximate a probability distribution than an arbitrary nonlinear function [Julier 96]. We propose to use a density interpolation technique to approximate the observation likelihood function by a GMM. Furthermore, to prevent the number of mixtures from growing exponentially over time, a density approximation based on the Expectation Maximization (EM) algorithm is applied, resulting in a compact GMM representation of the posterior pdf.

This chapter is structured as follows: Section 5.2 introduces the audio-visual state and measurement models. The proposed tracking framework is detailed in Section 5.3. Experimental results are evaluated in Section 5.4, and conclusions drawn in Section 5.6.

5.2 System Model

Consider a stream of synchronized sensory inputs, i.e., an image sequence and multi-channel microphone signals. Let t denote the time-step index of the audio-visual stream of data. The audio-visual tracking problem is modeled as a discrete-time Dynamic State Space Model (DSSM). The hidden system state, $\mathbf{X}_t \in \mathbb{R}^{\mathrm{D}}$, with initial distribution, $\mathrm{P}(\mathbf{X}_0)$, represents the two-dimensional (2D) location of N human talkers on the image plane at time step t. The system state evolves over time as an unobserved first-order Markov process according to the state transition model given by the conditional pdf, $\mathrm{P}(\mathbf{X}_t|\mathbf{X}_{t-1})$. The audio-visual observations at t, denoted by $\mathbf{Z}_t = \{\mathbf{Z}_t^a, \mathbf{Z}_t^v\}$, are conditionally independent from all other variables given the state $\mathbf{X}_t$ and are generated according to the pdf:

$$\mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t) = \mathrm{P}(\mathbf{Z}_t^a|\mathbf{X}_t)\mathrm{P}(\mathbf{Z}_t^v|\mathbf{X}_t), \tag{5.1}$$

where $\mathbf{Z}_t^a$ denotes the auditory observations extracted from the audio frame and $\mathbf{Z}_t^v$ denotes the visual observations extracted from the image at time-step t. The DSSM can also be written as a set of system equations representing the process and observation models as:

$$\mathbf{X}_t = \mathbf{f}(\mathbf{X}_{t-1}, \mathbf{V}_t) \qquad \text{(process model)} \tag{5.2}$$

$$\mathbf{Z}_t^a = \mathbf{h}_a(\mathbf{X}_t, \mathbf{U}_t^a) \qquad \text{(auditory observation model)} \tag{5.3}$$

$$\mathbf{Z}_t^v = \mathbf{h}_v(\mathbf{X}_t, \mathbf{U}_t^v) \qquad \text{(visual observation model)} \tag{5.4}$$

where $\mathbf{V}_t$ denotes the process noise that drives the dynamic system through a nonlinear state transition function, $\mathbf{f}$, and where $\mathbf{U}_t^a, \mathbf{U}_t^v$ denote the auditory and visual observation noise corrupting the observation of the system state through the nonlinear observation functions $\mathbf{h}_a$ and $\mathbf{h}_v$, respectively.

The face detector proposed in [Viola 04], and implemented in OpenCV, is used to obtain the visual observations $\mathbf{Z}_t^v$. However, the face detector is only reliable when a frontal face is presented and uninformative if a person turned his/her face away from the camera. The multi-source localization approach in [Evers 14] is used to obtain auditory observations, $\mathbf{Z}_t^a$, in the form of DOA estimates in azimuth and inclination.

A major challenge in audio-visual processing is the representation of auditory and visual observations in a common space. In the present work, a geometric transformation [Sanchez-Riera 12] is therefore applied to the DOAs in order to map the source directions from spherical space to a pixel position on the image plane. Therefore, auditory DOAs, visual facial detections, and the desired system states can be treated in the same mathematical space.

5.3 PROPOSED METHOD

Given all available observations $\mathbf{Z}_{1:t-1} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_{t-1}\}$ up to time-step $t-1$, the current posterior at t is predicted using the state transition model and prior, $\mathrm{P}(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})$, as:

$$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \int \mathrm{P}(\mathbf{X}_t|\mathbf{X}_{t-1})\mathrm{P}(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1})\,\mathrm{d}\mathbf{X}_{t-1}. \tag{5.5}$$

As the audio-visual observations, $\mathbf{Z}_t$, become available at t, the state can be updated using Bayes's theorem, i.e.,

$$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t}) = \frac{1}{C}\mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t)\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1}), \tag{5.6}$$

where

$$C = \mathrm{P}(\mathbf{Z}_t|\mathbf{Z}_{1:t-1}) = \int \mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t)\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1})\,\mathrm{d}\mathbf{X}_t \tag{5.7}$$

is a normalization constant. The posterior $\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t})$, also referred to as the filtering distribution, is used as the prior distribution at the next time-step $t+1$.

5.3.1 PREDICTED PDF

Assume that the prior pdf at t is given by a GMM with n_{t-1} components, i.e.,

$$\mathrm{P}(\mathbf{X}_{t-1}|\mathbf{Z}_{1:t-1}) = \sum_{i=1}^{n_{t-1}} \pi_{t-1}^i \mathcal{N}(\mathbf{X}_{t-1}; \mathbf{X}_{t-1}^i, \mathbf{\Sigma}_{t-1}^i), \tag{5.8}$$

where $\{\pi_{t-1}^i\}_{i=1}^{n_{t-1}}$ are the mixing weights with $\sum_{i=1}^{n_{t-1}} \pi_{t-1}^i = 1$, $\{\mathbf{X}_{t-1}^i\}_{i=1}^{n_{t-1}}$ are the mean vectors with $\mathbf{X}_{t-1}^i \in \mathbb{R}^{\mathrm{D}}$ and $\{\mathbf{\Sigma}_{t-1}^i\}_{i=1}^{n_{t-1}}$ are the covariance matrices with $\mathbf{\Sigma}_{t-1}^i \in \mathbb{R}^{\mathrm{D}\times\mathrm{D}}$. The UT enables the propagation of the means and covariance matrices through a nonlinear function *e.g.,* as in (5.2). To calculate the statistics of $\mathbf{X}_t$ that undergoes a nonlinear transformation, a set of $2\mathrm{D}+1$ weighted samples or sigma points, $\{\boldsymbol{S}_{(i,j)}\}_{j=0,i=1}^{2\mathrm{D},n_{t-1}}$, are carefully chosen so that they capture the mean and covariance of the system state. A selection scheme that satisfies this requirement is [Haykin 01]:

$$\begin{aligned}
&\boldsymbol{S}_{(i,0)} = \mathbf{X}_{t-1}^i, \qquad w_{(i,0)} = \tfrac{\lambda}{(\mathrm{D}+\lambda)}, && j = 0 \\
&\boldsymbol{S}_{(i,j)} = \mathbf{X}_{t-1}^i - \big(\sqrt{(\mathrm{D}+\lambda)\mathbf{\Sigma}_{t-1}^i}\big)_j, && j = 1, \ldots, \mathrm{D} \\
&\boldsymbol{S}_{(i,j)} = \mathbf{X}_{t-1}^i + \big(\sqrt{(\mathrm{D}+\lambda)\mathbf{\Sigma}_{t-1}^i}\big)_{j-\mathrm{D}}, && j = \mathrm{D}+1, \ldots, 2\mathrm{D} \\
&w_{(i,j)} = \tfrac{1}{2(D+\lambda)} \quad j = 1, \ldots, 2\mathrm{D},
\end{aligned} \tag{5.9}$$

where $w_{(i,j)}$ is the weight associated with the jth sigma point such that $\sum_{j=0}^{2\mathrm{D}} w_{(i,j)} = 1$ and λ is a scaling parameter and $\big(\sqrt{(\mathrm{D}+\lambda)\mathbf{\Sigma}_{t-1}^i}\big)_j$ is the j^{th} row of the matrix square root of $(\mathrm{D}+\lambda)\mathbf{\Sigma}_{t-1}^i$. As the random variable undergoes a non-linear transformation, these points are propagated through this non-linear function and are used to reconstruct the new means and covariance matrices. The estimated mean, $\bar{\mathbf{X}}_t^i$, and covariance, $\bar{\mathbf{\Sigma}}_t^i$, of the i^{th} Gaussian component in the predicted distribution are approximated using a weighted sample mean

and covariance of the sigma points. Finally, the GMM approximating the predicted pdf is given by:

$$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t-1}) = \sum_{i=1}^{n_{t-1}} \pi_t^i \mathcal{N}(\mathbf{X}_t; \bar{\mathbf{X}}_t^i, \bar{\mathbf{\Sigma}}_t^i), \tag{5.10}$$

where

$$\pi_t^i = \pi_{t-1}^i, \qquad \bar{\mathbf{X}}_t^i = \sum_{j=0}^{2\mathrm{D}} w_{(i,j)} \boldsymbol{S}_{(i,j)}, \quad \text{and}$$

$$\bar{\mathbf{\Sigma}}_t^i = \sum_{j=0}^{2\mathrm{D}} w_{(i,j)} \left(\boldsymbol{S}_{(i,j)} - \bar{\mathbf{X}}_t^i \right) \left(\boldsymbol{S}_{(i,j)} - \bar{\mathbf{X}}_t^i \right)^\top + \mathbf{Q}_t.$$

Here, $\mathbf{Q}_t$ is the process noise covariance matrix.

5.3.2 LIKELIHOOD

The observation likelihood is a measure for evaluating which configuration of $\mathbf{X}_t$ best matches the observations $\mathbf{Z}_t$ at time-step t. In order to efficiently explore the possible configuration space of $\mathbf{X}_t$ and obtain a good approximation for the posterior distribution, m samples $\{\mathbf{X}_i\}_{i=1}^m$, are generated from the Gaussian Mixture (GM) distribution given in (5.10). For each sample $\mathbf{X}_i$, the audio-visual likelihood l_i is computed as:

$$l_i = \beta_1 l_{app}(\mathbf{X}_i) + \beta_2 l_a(\mathbf{X}_i) + \beta_3 l_v(\mathbf{X}_i), \tag{5.11}$$

where l_{app} is function that evaluate appearance similarity between tracked targets and image regions indicated by the hypothesis state configuration $\mathbf{X}_i$. l_a is a function that measure the distance between $\mathbf{X}_i$ and auditory observations. l_v measures the distance between $\mathbf{X}_i$ and visual observations. $\{\beta_i\}_{i=1}^3$ are parameters that control the influence of the likelihood functions. We set a large value for β_1 since the appearance similarity is more reliable compared to other two distance measures. We note that the likelihood in (5.11) increases as audio-visual observations become available. However, the likelihood does not decrease for missing detections. This property makes the likelihood model robust to absence of audio-visual observations, *e.g.,*, in time of missed face detections and speech inactivity.

Furthermore, to obtain a continuous approximation of the likelihood given the discrete samples $\{\mathbf{X}_i\}_{i=1}^m$ and associated likelihood $\{l_i\}_{i=1}^m$, the Radial Basis Function (RBF) [Poggio 89] is used for interpolation. Therefore, a Gaussian kernel is assigned to each sample $i = 1, \ldots, m$, such that the likelihood of $\mathbf{X}_j$ induced by the i^{th} kernel is given by

$$\mathrm{P}_i(\mathbf{X}_j) = \mathcal{N}(\mathbf{X}_j; \mathbf{X}_i, \mathbf{P}_i), \tag{5.12}$$

where the sample location, $\mathbf{X}_i$, is used as the mean and the covariance, or the kernel bandwidth, $\mathbf{P}_i$, is set to k-nearest neighbors (KNN) distance, i.e.,[1]

$$\mathbf{P}_i = c \, \mathrm{diag}\big(\mathrm{KNN}_1^i(k), \ldots, \mathrm{KNN}_\mathrm{D}^i(k)\big)\mathbf{I}, \tag{5.13}$$

[1]The approach can be naturally extended to more complex approaches to kernel bandwidth selection, discussed in, e.g., [Sheather 91].

where c is a constant that depends on the number of samples and the dimensionality, $\mathbf{I}$ is the D-dimensional identity matrix, and $\mathrm{KNN}_j^i(k)$ is the KNN distance of sample i in the j^{th} dimension. Therefore, the observation likelihood $\mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t)$ is approximated by $n_t \ll m$ Gaussians, such that

$$\mathrm{P}(\mathbf{Z}_t|\mathbf{X}_t) = \sum_{i=1}^{n_t} w_i \mathcal{N}(\mathbf{X}_t; \mathbf{X}_t^i, \mathbf{P}_t^i). \tag{5.14}$$

The weight, w_i, of kernel $i = 1, \ldots, n_t$, is computed by solving the constrained non-negative least square problem [Poggio 89]:

$$\begin{gathered} \arg\min_{\mathbf{w}} \ \|\mathbf{A}\mathbf{w} - \mathbf{b}\|^2 \\ \text{subject to elements of } \mathbf{w} \geq 0 \end{gathered} \tag{5.15}$$

where $\mathbf{A} \in \mathbb{R}^{m \times m}$ is a design matrix with each element (i, j) given by $\mathrm{P}_i(\mathbf{X}_j)$, the matrix $\mathbf{b} \in \mathbb{R}^{m \times 1}$ contains l_i for each row $i = 1, \ldots, m$, and $\mathbf{w} = [w_1, w_2, \ldots, w_m]^\top$ is the kernel weight column-vector.

5.3.3 POSTERIOR PDF

Both the predicted density in (5.10) and the observation likelihood in (5.14) are represented by GMs. Therefore, the posterior pdf in (5.6) is equivalent to a product of two GMs, and is therefore equivalent to a GM with an exponentially increasing number of components, i.e.,

$$\begin{aligned} &\left(\sum_{i=1}^{n_{t-1}} \pi_t^i \mathcal{N}(\mathbf{X}_t; \bar{\mathbf{X}}_t^i, \bar{\mathbf{\Sigma}}_t^i) \right) \left(\sum_{j=1}^{n_t} \tau_t^j \mathcal{N}(\mathbf{X}_t; \mathbf{X}_t^j, \mathbf{P}_t^j) \right) \\ &\qquad = \sum_{i=1}^{n_{t-1}} \sum_{j=1}^{n_t} w_t^{ij} \mathcal{N}(\boldsymbol{\mu}_t^{ij}, \mathbf{\Sigma}_t^{ij}), \end{aligned} \tag{5.16}$$

where

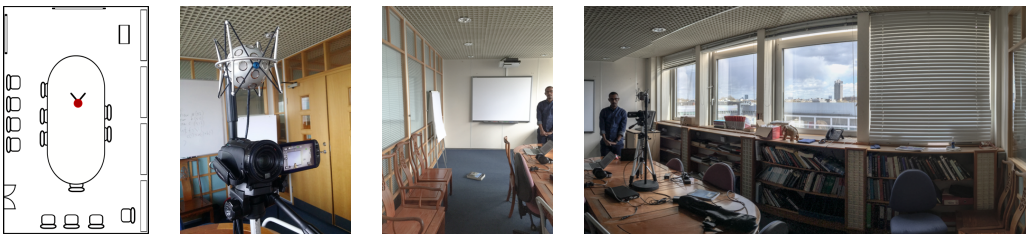
$$w_t^{ij} = \frac{\pi_t^i \tau_t^j \exp\left(-\frac{1}{2} (\mathbf{X}_t^j - \bar{\mathbf{X}}_t^i)^\top \left(\bar{\mathbf{\Sigma}}_t^i + \mathbf{P}_t^j \right)^{-1} (\mathbf{X}_t^j - \bar{\mathbf{X}}_t^i) \right)}{(2\pi)^{(\mathrm{D}/2)} |\bar{\mathbf{\Sigma}}_t^i + \mathbf{P}_t^j|^{1/2}} \tag{5.17}$$

$$\mathbf{\Sigma}_t^{ij} = \left(\left(\bar{\mathbf{\Sigma}}_t^i \right)^{-1} + \left(\mathbf{P}_t^j \right)^{-1} \right)^{-1} \tag{5.18}$$

$$\boldsymbol{\mu}_t^{ij} = \mathbf{\Sigma}_t^{ij} \left(\left(\bar{\mathbf{\Sigma}}_t^i \right)^{-1} \bar{\mathbf{X}}_t^i + \left(\mathbf{P}_t^j \right)^{-1} \mathbf{X}_t^j \right). \tag{5.19}$$

The resulting density function in (5.16) is a weighted mixture of Gaussians with $n_{t-1} \times n_t$ components.

To mitigate an exponential explosion in the number of components, we first note that typically many of the kernels correspond to weights close to zero and are therefore negli-



(a) (b) (c) (d)

Figure 5.1: Camera-microphone and recording room setup. (a) Recording room schematic (the position of the camera-microphone pair is shown with a red dot. (b) The camera-microphone setup used to recording test scenarios. (c) and (d) photos taken inside the recording room.

gible. To remove stochastically irrelevant kernels, the weighted data EM in [Gebru 16a] is used to fit a GMM to (5.16). The optimal number of GM components, m_t, for fitting is selected in a principled manner based on the information-theoretic Minimum Message Length (MML) principle. We note that although the weighted data EM is not a general density approximation method, it does preserve significant mode locations and approximates well-separated GMs accurately. Thus, the final posterior distribution at time-step t is given by:

$$\mathrm{P}(\mathbf{X}_t|\mathbf{Z}_{1:t}) = \sum_{i=1}^{m_t} \pi_t^i \mathcal{N}(\mathbf{X}_t; \mathbf{X}_t^i, \mathbf{\Sigma}_t^i), \tag{5.20}$$

where m_t is the number of components [Gebru 16a].

5.4 Experimental Setup

The camera-microphone setup shown in Fig. 5.1 is used to gather audio-visual recordings in order to evaluate the performance of the proposed model. The setup consists of a HD video camera and a 32 channel spherical microphone array (Eigenmike). This microphone is particularly suitable for capturing all three dimensions of the acoustic environment, including both ambient sounds and sounds from particular directions. The camera provides video frames with a resolution of 1920×1200 pixels at a rate of 25 Frames-Per-Second (FPS). The original audio signals are captured at 48 kHz, and are downsampled to 16 kHz. The acoustic impulse generated by a "clapperboard" is used for audio-visual synchronization.

Two scenarios are recorded in order to test the robustness of the proposed tracking model to dynamic scenes. The scenarios are referred to in the following as **S1** and **S2**. Both recordings are taken in the meeting room illustrated in Fig. 5.1. In both scenarios, two talkers are speaking whilst moving within the room. The talkers occasionally occlude each other and move in and out of the camera field of view. In **S1**, the participants take speech turns, whilst the talkers speak simultaneously most of the time in **S2**. Scenario **S1** has 3000 video frames (2 minutes); **S2** has 2000 video frames (1 minute and 20 seconds). For both scenarios, the ground truth is identified by hand-annotating in each video frame the talker positions with bounding boxes around faces, 2D locations inferred from the bounding boxes, and target ID.

Table 5.1: Tracking results comparison. ↑ denotes higher scores indicate better results, and ↓ denotes lower scores indicate better results.

Sequence	Methods	MOTA (in %)↑	MOTP (in %)↑	OMAT↓	OSPA↓
S1	P-AV	83.6	89.1	186.6	112.7
	P-V	72.3	82.8	245.6	234.0
	OV[Ba 16]	45.7	56.4	378.6	367.8
S2	P-AV	89.8	84.8	201.4	198.0
	P-V	86.3	80.9	200.4	193.6
	OV[Ba 16]	44.3	47.6	356.7	367.8

Face detection yields fully automatic initialization. The algorithm initializes a new track for a person that has subsequent detections with overlapping bounding boxes, which are neither occluded nor associated to an already existing tracks. The initial mean positions are drawn from a Gaussian distribution centered at the detection center. The initial size corresponds to the detection size. An appearance model based on RGB color histogram is initialized from the image region around the initial bouding box. Conversely, if not enough detections are found for the same target within T consecutive frames, the track is automatically terminated. Since addition and termination of targets, respectively, increases and decreases the system state dimension, this procedure is performed at the start of each-time step, prior to density approximation. The Hungarian algorithm [Kuhn 55] is used to obtain the optimal face detection to existing target association. The association cost matrix is based on the pairwise similarity of the RGB histogram between existing targets and detections.

The common CLEAR MOT metrics [Bernardin 08] consisting of multiple metrics are used for evaluation. The Multiple Object Tracking Precision (MOTP) evaluates the intersection area over the union area of bounding boxes. The Multiple Object Tracking Accuracy (MOTA) calculates the accuracy composed of false negatives, false positives and identity switching. Furthermore, the Optimal MAss Transfer (OMAT) metric [Hoffman 04] and the Optimal Aub-Pattern Assignment (OSPA) metric [Schuhmacher 08] are used in order to evaluate tracking performance independent of track labelling. These metrics compute set distances between the ground truth set of objects present in the scene and the set of objects estimated by the tracker.

5.5 EXPERIMENTAL VALIDATION

The performance of the proposed tracking model using auditory and visual observations (P-AV) is compared against the proposed tracking model without the auditory data (P-V) and the multi-person visual tracker (OV) in [Ba 16].

The results are summarized in Table 5.1, whilst the audio-visual tracking results for the proposed P-AV are illustrated for a few video frames in Fig. 5.2, Fig. 5.3, Fig. 5.4 and

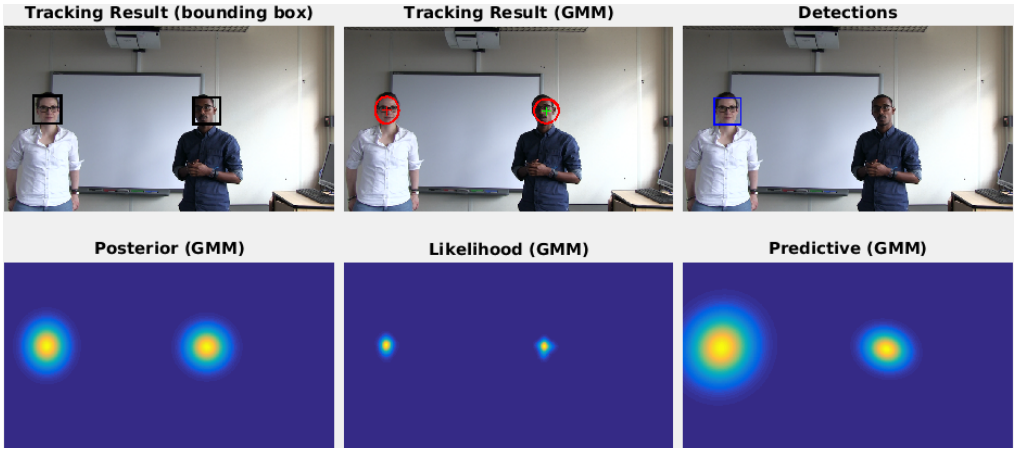


Figure 5.2: Results obtained on scenario **S1** at video frame #351. (top-left): Bounding boxes showing the final tracking results, width and height are calculated from 95% confidence region of the posterior Gaussian, (top-middle): Plots of the final posterior GMM: (top-right): Face detections used as visual observations, (bottom row): 2D density plot of the posterior, likelihood and predictive GMM.

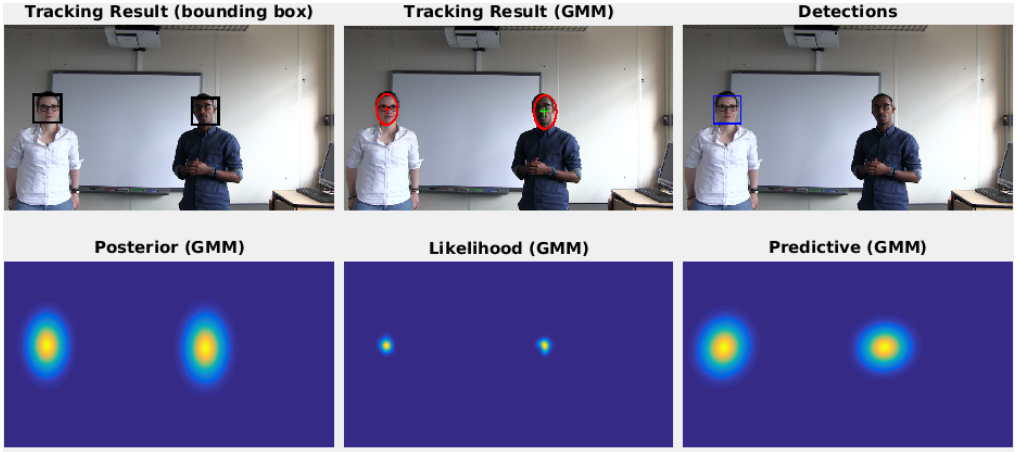


Figure 5.3: Results obtained on scenario **S1** at video frame #352.

Fig. 5.5. The full videos, Matlab code and additional examples are available online[2].

The comparison between P-AV and P-V highlights that constructive exploitation of the auditory observations leads to improved disambiguation of the states of the moving talkers. An improvement of 11.3 percentage points in MOTA is achieved using P-AV compared to P-V for S1. For S2, the improvement for the audio-visual tracker corresponds to 3.5 percentage points. The difference in results between S1 and S2 can be explained by the difficulty of the scenarios. In S2, the position of the talkers is mostly static, mainly affected by body and head rotations. Therefore, as facial occlusions are rare, face detection is highly reliable. In S1, frequent visual occlusions and obscurations as well as crossing speaker paths lead to less reliable visual detections. Furthermore, the talkers frequently speak simultaneously. Hence, audio DOAs for both talkers are exploited constructively

[2] `http://team.inria.fr/perception/avtracking_by_dabf/`

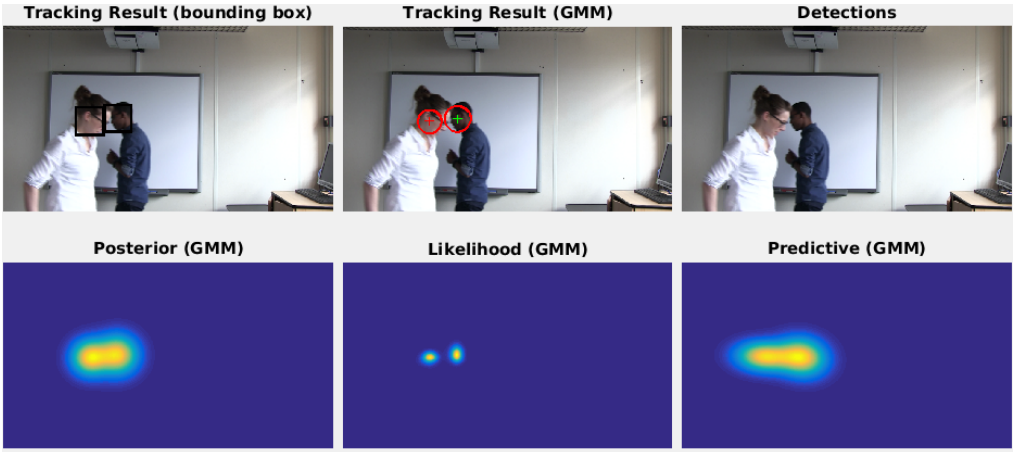


Figure 5.4: Results obtained on scenario **S1** at video frame #451.

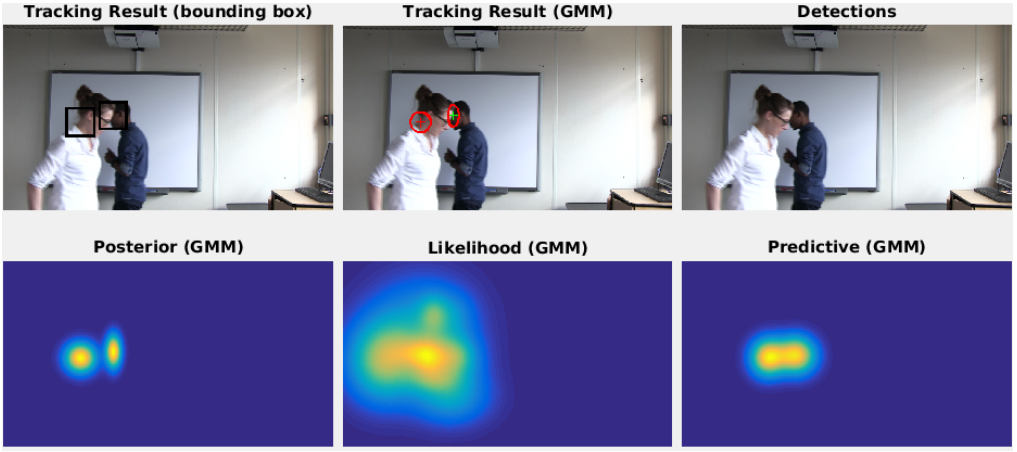


Figure 5.5: Results obtained on scenario **S1** at video frame #452.

for audio-visual disambiguation in frames where faces are visually occluded.

The results also highlight that the proposed approach outperforms the benchmark OV in both scenarios. This is due to propagation of multiple Gaussians for each target at each time-step by the proposed model, leading to increased robustness to measurement ambiguity. For S1, P-AV therefore results in an improvement of 37.9 and 32.7 percentage points in MOTA and MOTP respectively compared to OV. For S2, an improvement of 45.5 and 32.7 percentage points is achieved.

5.6 Conclusions

This chapter presented a novel approach to audio-visual tracking that jointly utilizes multiple DOAs obtained by sound source localization and facial detections to estimate the trajectories of multiple people in pixel space. Within a Bayesian framework, the posterior pdf is propagated using a GMM. To retain the Gaussianity over time, the non-linear observation likelihood function is approximated by a density interpolation technique. Fur-

thermore, to avoid the exponential explosion of Gaussians over time, an EM algorithm is used to cluster Gaussians, thereby retaining only statistically relevant components. Results based on measurements with a camcorder and eigenmike demonstrate significant improvements in performance for audio-visual tracking compared to using only visual data.

CHAPTER 6

TRACKING THE ACTIVE SPEAKER BASED ON AV DATA

Any multi-party conversation system benefits from speaker diarization, that is, the assignment of speech signals among the participants. We here cast the diarization problem into a speaker tracking formulation whereby the active speaker is detected and tracked over time. A probabilistic model exploits the on-image (spatial) coincidence of visual and auditory observations and infers a single latent variable which represents the identity of the active speaker. Both visual and auditory observations are explained by a recently proposed weighted-data mixture model, while several options for the speaking turns dynamics are fulfilled by a multi-case transition model. The modules that translate raw audio and visual data into on-image observations are also described in detail. The performance of the proposed tracker is tested on challenging data-sets that are available from recent contributions which are used as baselines for comparison.

6.1 INTRODUCTION

In human-computer interaction (HCI) and human-robot interaction (HRI) it is often necessary to solve multi-party dialog problems. For example, if two or more persons are engaged in a conversation, one important task to be solved, prior to automatic speech recognition (ASR) and natural language processing (NLP), is to correctly assign speech segments to corresponding speakers. This problem is often referred to as speaker diarization in the speech/language processing literature and a number of solutions has been recently proposed, e.g., [Anguera Miro 12]. When only auditory data are available, the task is very difficult because of the inherent ambiguity of mixed acoustic signals captured by the microphones. An interesting alternative consists in combining auditory and visual data. This is motivated by the fact that humans understand complex auditory and visual information, and often uses one to disambiguate the other. The two modalities provide complementary information and hence audio-visual approaches to speaker diarization are likely to be more robust than audio-only approaches.

Several audio-visual diarization methods were recently proposed, *e.g.*,, [Noulas 12, Anguera Miro 12, Minotto 15]. Noulas et al. [Noulas 12] proposed a graphical model, where latent discrete variables represent speaker identities and speaker visibilities over time. The main limitation of [Noulas 12] as well as of other audio-visual approaches reviewed in [Anguera Miro 12] is that these methods require the detection of frontal faces and of mouth/lip motions. Indeed, audio-visual association is often solved using the temporal correlation, over several seconds, between facial features and audio features [Potamianos 03]. Minotto et al. [Minotto 15] learn an SVM classifier using labeled audio-visual features, which is dependent on the acoustic properties of the training data. They combine voice activity detection with sound-source localization using a linear microphone array. The latter can only provide the azimuth (horizontal) sound direction. Their method relies on mouth tracking, hence frontal views of the speakers are required as well.

More generally, audio-visual association for speaker diarization can be achieved on the premise that a speech signal *coincides* with a person that is visible and that emits a sound. This coincidence must occur both in space and time. In formal multi-party conversations, diarization is facilitated by participants that talk sequentially, presence of a short silence between speech turns, and participants facing the cameras while remaining seated or static. In these cases, audio-visual association based on temporal coincidence seems to provide satisfactory results, e.g., [Kidron 07]. In informal settings which are very common, particularly in HRI, the situation is much more complex. The perceived audio signals are corrupted by environmental noise, reverberations, and several persons may occasionally speak simultaneously. Moreover, people may wander around, turn their heads away from the sensors, be occluded by other people, suddenly disappear from the camera field of view, and appear again later on.

These problems were addressed by several authors in different ways. For example, [Gatica-Perez 07] proposed a multi-speaker tracker using approximate inference implemented with a Markov chain Monte Carlo particle filter (MCMC-PF). In [Naqvi 10a] a 3D visual tracker is proposed, based on MCMC-PF as well, to estimate the positions and velocities of the participants which are then passed to blind source separation based on beamforming [Van Veen 88]. Reported experiments of both [Gatica-Perez 07, Naqvi 10a] require a network of distributed cameras to guarantee that frontal views of the speakers are always available. More recently, [Kilic 15a] proposed to use audio information to assist the particle propagation process and to weight the observation model. This implies that audio data are always available and that they are reliable enough to properly relocate the particles. While audio-visual multiple persons tracking methods provide an interesting methodology, they do not address the challenging speaker diarization problem.

In this work, we propose to enforce audio-visual spatial coincidence [Khalidov 11a, Alameda-Pineda 15a], rather than temporal coincidence [Kidron 07, Sargin 07], into diarization. We consider a setup consisting of people that are engaged in a multi-party conversation while they are free to move and to turn their attention away from the cameras. We propose to combine an online multi-person visual tracker [Bae 14], with a voice activity detector [Sohn 99], and a sound-source localizer [Deleforge 15b], e.g., Fig. 4.2.

Assuming that the image and audio sequences are synchronized, we propose to group auditory features and visual features based on the premise that they share a common location if they are generated by the same speaker. We introduce a latent variable representing the active-speaker, and we devise an on-line tracker such that the identity and location of the active speaker is estimated over time. We propose a generative observation model, based on the recently proposed weighted-data Gaussian mixture [Gebru 16b], that evaluates the posterior probability of an observed person to be the active speaker, conditioned by the output of a multi-person visual tracker, a sound-source localizer, and a voice activity detector. We also propose a dynamic model that allows to estimate the active speaker using temporal transition probabilities modeling speaking activity transition priors from frame $t-1$ to frame t. The proposed on-line tracking method uses an efficient exact inference algorithm.

The remainder of this chapter is organized as follows. Section 6.2 formally describes the proposed exact inference method; section 6.2.1 describes the audio-visual generative observation model; section 6.2.2 describes the proposed transition probabilities model. Section 6.3 describes implementation details and experiments. Finally, section 6.4 draws some conclusions. Videos, Matlab code and additional examples are available online.[1]

6.2 Model for Tracking the Active Speaker

We start by introducing some notations and definitions. Upper-case letters denote random variables while lower-case letters denote their realizations. We consider an image sequence that is synchronized with an audio sequence and let t denote the frame index of both visual and audio modalities (without loss of generality, one can assume that audio and visual frames have the same temporal length). Let N be the maximum number of visual observations at any time. Hence at frame t we have $\mathbf{X}_t = (\boldsymbol{X}_{t1}, \ldots, \boldsymbol{X}_{tn}, \ldots, \boldsymbol{X}_{tN}) \in \mathbb{R}^{2\times N}$, where the random variable $\boldsymbol{X}_{tn}$ corresponds to the location of person n at t. We also introduce the binary variables $\boldsymbol{V}_t = (V_{t1}, \ldots, V_{tN})$ such that $V_{tn} = 1$ if person n is detected *visible* in frame t and $V_{tn} = 0$ if the person is not detected. The time series $\mathbf{X}_{1:t} = \{\mathbf{X}_1, \ldots, \mathbf{X}_t\}$ and associated visual presence masks $\mathbf{V}_{1:t} = \{V_1, \ldots, V_t\}$ can be estimated using a multi-person tracker. We perform multi-person tracking using [Bae 14] (see section 6.3 below). Let $N_t = \sum_n V_{tn}$ denote the number of persons that are visible at t.

We also consider auditory information. Audio activity is described by the binary variable $A_t \in \{0, 1\}$ that is estimated using voice activity detection (VAD) and which is equal to 1 if audio activity is detected at t and 0 otherwise. Whenever a frame has audio activity, a binaural (two microphones) sound-source localization (SSL) algorithm provides spatial audio information: a sound-source direction (azimuth and elevation) is mapped onto the image plane, e.g., [Deleforge 15b], Fig. 4.2, and section 6.3 below.

Let K be the number of sound-source directions estimated at frame t when $A_t = 1$. Let $\mathbf{Y}_t = (\boldsymbol{Y}_{t1}, \ldots, \boldsymbol{Y}_{tk}, \ldots, \boldsymbol{Y}_{tK}) \in \mathbb{R}^{2\times K}$ denote the K sound-source directions at

[1] https://team.inria.fr/perception/avdiarization/

t. Hence, VAD combined with SSL estimate a time series of sound locations $\mathbf{Y}_{1:t} = \{\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_t\}$ and associated *audio-activity binary masks* $\mathbf{A}_{1:t} = \{A_1, \ldots, A_t\}$.

The objective is to track the active speaker which amounts to associate over time the audio activity (if any) with one of the tracked persons. This is also referred to as audio-visual speaker diarization, e.g., [Noulas 12] which is addressed below in the framework of temporal graphical models; A time-series of discrete latent variables is introduced, $\mathbf{S}_{1:t} = \{S_1, \ldots, S_t\}$ such that $S_t = n, n \in \{1, \ldots, N\}$ if person n is both observed and speaks at t, and $S_t = 0$ if none of the visible persons speaks at t. Notice that $S_t = 0$ encompasses two cases, namely (i) there is audio activity at t ($A_t = 1$) but sound-source locations cannot be associated with one of the visible persons, and (ii) there is no audio activity at t ($A_t = 0$). The active-speaker tracking can be formulated as a maximum a posteriori (MAP) estimation problem:

$$\hat{s}_t = \underset{s_t}{\operatorname{argmax}} \, \mathrm{P}(S_t = s_t | \boldsymbol{x}_{1:t}, \boldsymbol{v}_{1:t}, \boldsymbol{y}_{1:t}, \boldsymbol{a}_{1:t}). \tag{6.1}$$

The posterior probability (6.1) can be written as:

$$\mathrm{P}(S_t = s_t | \boldsymbol{u}_{1:t}) = \frac{\mathrm{P}(\boldsymbol{u}_t | S_t = s_t, \boldsymbol{u}_{1:t-1}) \mathrm{P}(S_t = s_t | \boldsymbol{u}_{1:t-1})}{\mathrm{P}(\boldsymbol{u}_t | \boldsymbol{u}_{1:t-1})}, \tag{6.2}$$

where we used the notation $\boldsymbol{u}_t = (\boldsymbol{x}_t, \boldsymbol{v}_t, \boldsymbol{y}_t, a_t)$. The numerator of (6.2) expands as:

$$= \mathrm{P}(\boldsymbol{u}_t | S_t = s_t) \sum_{i=0}^{N} \mathrm{P}(S_t = s_t | S_{t-1} = i) \mathrm{P}(S_{t-1} = i | \boldsymbol{u}_{1:t-1}). \tag{6.3}$$

The denominator of (6.2) expands as:

$$= \sum_{j=0}^{N} \Big(\mathrm{P}(\boldsymbol{u}_t | S_t = j) \Big(\sum_{i=0}^{N} \mathrm{P}(S_t = j | S_{t-1} = i) \times \mathrm{P}(S_{t-1} = i | \boldsymbol{u}_{1:t-1}) \Big) \Big). \tag{6.4}$$

The evaluation of this recursive relationship requires (i) the joint audio-visual likelihood $\mathrm{P}(\boldsymbol{u}_t | S_t = s_t)$, (ii) the transition probabilities $\mathrm{P}(S_t = j | S_{t-1} = i)$, and (iii) the initial posteriors $\mathrm{P}(S_1 = s_1 | \boldsymbol{u}_1), s_1 \in \{0, 1, \ldots, n, \ldots, N\}$. We note that because N is small (of the order of 10), the exact evaluation of (6.1) is tractable and hence solving the MAP problem (6.2) is straightforward.

6.2.1 The Audio-Visual Association Model

In this section we derive an expression for the joint audio-visual likelihood. One crucial feature of the proposed model is its ability to robustly associate the acoustic activity at frame t with a person. The generative model that is proposed below assigns the audio activity, if any, to a person, or to nobody. In this context, let Z_{tk} be the (audio) observation-to-person assignment variable in our mixture model. The case $A_t = 1$ is first considered, namely there is audio activity at t. The source location observed variables $\boldsymbol{Y}_{tk}$ are assumed to be drawn from the following WD-GMM (weighted-data Gaussian

mixture model) [Gebru 16b]:

$$\mathrm{P}(\boldsymbol{y}_{tk}|\boldsymbol{x}_t, \boldsymbol{v}_t, A_t = 1; \boldsymbol{\theta}_t, \boldsymbol{\phi}_{tk}) = \sum_{n=1}^{N} \pi_{tn} v_{tn} \mathcal{N}(\boldsymbol{y}_{tk}|\boldsymbol{x}_{tn}, \frac{1}{w_{tk}}\boldsymbol{\Sigma}_{tn}), \tag{6.5}$$

where $\boldsymbol{\theta}_t = (\{\pi_{tn}\}_{n=1}^N, \{\boldsymbol{\Sigma}_{tn}\}_{n=1}^N)$ denotes the GMM free parameters, namely the priors $\pi_{tn} = \mathrm{P}(S_t = n)$, $\sum_{n=1}^{N} v_{tn}\pi_{tn} = 1$ and the 2×2 covariance matrices $\boldsymbol{\Sigma}_{tn}$. In the proposed formulation, the mixture mean vectors, $\{\boldsymbol{x}_{tn}\}_{n=1}^N$ are observed and they correspond to image locations of people heads, while the visibility variables $\{v_{tn}\}_{n=1}^N$ allow to consider only those that are visible at t. For convenience we only address the case $N_t \geq 1$. Notice that this model comprises a weight variable $w_{tk} > 0$ drawn from a gamma distribution $\mathcal{G}(w; \alpha, \beta) = \Gamma^{-1}(\alpha)\beta^{\alpha} w^{\alpha-1} e^{-\beta w}$ with parameters $\boldsymbol{\phi} = (\alpha, \beta)$. There is a weight *associated with each audio observation* $\boldsymbol{y}_{tk}$ and one may notice that the weight acts as a precision, higher the weight more relevant the observation, and that the observed data are independent but not identically distributed.

The posterior probability of a sound-source direction to be associated with the n-th visible person writes [Gebru 16b]:

$$\begin{aligned}\eta_{tkn} &= \mathrm{P}(Z_{tk} = n|\boldsymbol{y}_{tk}, \boldsymbol{x}_t, \boldsymbol{v}_t, A_t = 1; \boldsymbol{\theta}_t, \boldsymbol{\phi}_{tk}) \\ &\propto \pi_{tn} v_{tn} \mathcal{P}(\boldsymbol{y}_{tk}|\boldsymbol{x}_{tn}, \boldsymbol{\Sigma}_{tn}, \alpha_{tk}, \beta_{tk}),\end{aligned} \tag{6.6}$$

where $\mathcal{P}$ denotes the Pearson type VII probability distribution function (the reader is referred to [Sun 10] for a recent discussion regarding this distribution, also called the Arellano-Valle and Bolfarine generalized t-distribution [Kotz 04]):

$$\mathcal{P}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\Sigma}, \alpha, \beta) = \frac{\Gamma(\alpha + d/2)}{|\boldsymbol{\Sigma}|^{1/2}\,\Gamma(\alpha)\,(2\pi\beta)^{d/2}} \left(1 + \frac{\|\boldsymbol{y} - \boldsymbol{x}\|_{\boldsymbol{\Sigma}}^2}{2\beta}\right)^{-(\alpha + \frac{d}{2})} \tag{6.7}$$

The WD-GMM formulation allows one to write the posterior distribution of w_{tk}, which is a gamma distribution because it is the conjugate prior of the precision of the Gaussian distribution:

$$\mathrm{P}(w_{tk}|Z_{tk} = n, \boldsymbol{y}_{tk}, \boldsymbol{x}_{tn}; \boldsymbol{\theta}_t, \gamma_{tk}, \delta_{tkn}) \propto \mathcal{G}(w_{tk}; \gamma_{tk}, \delta_{tkn}), \tag{6.8}$$

where $\gamma_{tk} = \alpha_{tk} + d/2$ and $\delta_{tkn} = \beta_{tk} + 1/2\,\|\boldsymbol{y}_{tk} - \boldsymbol{x}_{tn}\|_{\boldsymbol{\Sigma}_{tn}}^2$. This allows to evaluate (6.6), as well as the posterior mean of w_{tk}, namely:

$$\overline{w}_{tk} = \sum_{n=1}^{N} v_{tn}\eta_{tkn}\overline{w}_{tkn}, \tag{6.9}$$

where $\overline{w}_{tkn} = \gamma_{tk}/\delta_{tkn}$ is the conditional mean, which is needed to update the mixture parameters (proportions and covariances in our case) during the maximization step (please consult Section 3.5 and Section 5 in [Gebru 16b] for more details). By inspection of the above equations it is easily seen that the value of $\overline{w}_{tk}$ is small if the distances between an audio observation $\boldsymbol{y}_{tk}$ and the cluster centers $\boldsymbol{x}_{tn}$ are large. In other words, the weight associated with an observed sound location that is far away from the observed persons is small compared with the weight of an observed sound location that coincides with a

person location. Hence, the estimated value of w_{tk}, namely $\overline{w}_{tk}$, reduces the influence of outliers. Notice that the weights w_{tk} play a different role than the responsibilities η_{tkn}. Indeed, the responsibilities are normalized, $\sum_{n=1}^{N}\eta_{tkn}=1$, hence they can only account for a relative measure of the data relevance. Therefore, we use the estimated weights $\{\overline{w}_{tk}\}_{k=1}^{K}$ and an inlier/outlier threshold w_s to classify the audio observations into an inlier set $\mathcal{Y}_{\text{in}}$ and an outlier set $\mathcal{Y}_{\text{out}}$.

Altogether, this formulation allows one to characterize the audio activity of each observed person. Assuming that the audio observations are independent, one obtains the likelihood of person n to be the active speaker:

$$\mathrm{P}(\boldsymbol{y}_t,\boldsymbol{x}_t,\boldsymbol{v}_t,A_t=1|S_t=n)\propto\begin{cases}\sum\limits_{k\in\mathcal{Y}_{\text{in}}}\eta_{tkn}, & 1\leq n\leq N,\\ \sum\limits_{k\in\mathcal{Y}_{\text{out}}}\eta_{tkn}, & n=0.\end{cases}\tag{6.10}$$

If there is no audio activity at time t, $A_t=0$, then $S_t=0$ (there is no active speaker) and the likelihood of an active speaker is a uniform distribution:

$$\begin{aligned}\mathrm{P}(\boldsymbol{y}_t,\boldsymbol{x}_t,\boldsymbol{v}_t,A_t=0|S_t=n)&\propto\mathrm{P}(S_t=n|\boldsymbol{y}_t,\boldsymbol{x}_t,\boldsymbol{v}_t,A_t=0)\\&=\begin{cases}r & n=0\\ \frac{1-r}{N_t} & 1\leq n\leq N\end{cases}\end{aligned}\tag{6.11}$$

where $r\in[0,1]$ describes the probability that there is no audio activity at time t, i.e., either there is no visible person or none of the visible persons speaks.

6.2.2 The State Transition Model

The state transition probabilities, $\mathrm{P}(\boldsymbol{S}_t=j|\boldsymbol{S}_{t-1}=i)$, provide a temporal model for tracking speech turns. Several cases need to be considered based on the presence/absence of persons and on their speaking status (for convenience and without loss of generality we set $v_{t0}=1$):

$$\mathrm{P}(\boldsymbol{S}_t=j|\boldsymbol{S}_{t-1}=i)=\begin{cases}p_s & \text{if}\quad i=j\text{ and }v_{t-1i}=v_{ti}=1\\ (1-p_s)/N_t & \text{if}\quad i\neq j\text{ and }v_{t-1i}=v_{tj}=1\\ 0 & \text{if}\quad v_{t-1i}=v_{t-1j}=1\text{ and }v_{tj}=0\\ 1/N_t & \text{if}\quad v_{t-1i}=1,v_{ti}=0\text{ and }v_{tj}=1\\ 1/N & \text{if}\quad v_{t-1i}=0\text{ and }v_{tj}=0.\end{cases}\tag{6.12}$$

The first case of (6.12) defines the self-transition probability, p_s, e.g., $p_s=0.8$, of person i present at both $t-1$ and t. The second case defines the transition probability from person i present at $t-1$ to another person j present at t. The third case simply forbids transitions from person i present at $t-1$ to person j present at $t-1$ but not present at t. The fourth case defines the transition probability from person i present at $t-1$ but not present at t, to a person j present at t. The fifth case defines the transition probability from person i not present at $t-1$ to person j that is not present at t. These latter transition probabilities are only defined for completeness as transition between non-visible persons are forbidden by the observation model.

These five cases can be grouped in a compact way to yield the state transition proba-

bility matrix ($\delta_{ij} = 1$ if $i = j$ and 0 otherwise):

$$\mathrm{P}(\boldsymbol{S}_t = j|\boldsymbol{S}_{t-1} = i) = \frac{1 - v_{ti}}{N_t} + v_{t-1i}v_{tj} \times \left(p_s\delta_{ij} + \frac{(1 - p_s)(1 - \delta_{ij})}{N_t} + \frac{1 - v_{ti}}{N_t} \right). \tag{6.13}$$

One may easily verify that $\sum_{j=0}^{N} \mathrm{P}(\boldsymbol{S}_t = j|\boldsymbol{S}_{t-1} = i) = 1$.

6.3 Implementation and Experiments

As already outlined, the proposed active-speaker tracker may well be viewed as a diarization process summarized as follows: track multiple persons based on visual information, estimate the auditory activity, and associate this activity to one of the tracked persons. Unlike existing audio-visual diarization approaches, which assume that the participants are always facing the cameras, the proposed model can deal with participants that are temporarily occluded, or who come in and out of the field of view of the camera. Unfortunately there are no publicly available datasets that include participants that take speech turns while they wander around, occlude each other and move in and out of the camera field of view.

Therefore we recorded our own data,[2] gathered with two microphones and one camera e.g., Fig. 4.2. The audio data are delivered by two microphones plugged into the ears of an acoustic dummy head; the visual data are delivered by a video camera. The two modalities are synchronized such that the video frames are temporally aligned with the audio samples. The videos are recorded at 25 FPS while the audio signals are sampled at 48000 Hz. With this setup, we gathered two scenarios, the *counting* scenario, Fig. 6.1 and the *chat* scenario, Fig. 6.2. The *counting* sequence has 500 video frames (20 seconds) while the *chat* sequence has 850 video frames (34 seconds).

We briefly describe the multi-person tracking and sound-source localization techniques used to obtain estimates of our observed auditory and visual variables (Section 6.2.1). Among the visual tracking methods that are currently available, we chose the multi-person tracker of [Bae 14]. This method has several advantages, namely (i) it robustly handles fragmented tracks, which are due to occlusions or to unreliable detections, and (ii) it performs online discriminative learning to handle similar appearances of different persons. The multi-person tracker provides realizations of the visual observation variables $\mathbf{X}_{1:t}$ and associated *visual-presence binary masks* $\mathbf{V}_{1:t}$, as explained in detail in Section 6.2.

We adopted the sound-source localization method of [Deleforge 15b] to estimate sound directions with two degrees of freedom (azimuth and elevation). A prominent advantage of this method, in the context of audio-visual analysis, is that it provides a built-in mechanism for mapping sound directions onto image locations. Hence sound-source directions are eventually expressed in pixel coordinates. In practice, the signals delivered by the two microphones are transformed in the Fourier domain in the following way: the short-time Fourier transform (STFT) is applied to a 0.064 s window of the two signals and

[2] `https://team.inria.fr/perception/avtrack1/`

this window is shifted along the time axis with 0.008 s hops (or 0.056 s overlap between successive windows). With this Fourier domain sampling, there are 5 feature vectors associated with each video frame. In order to increase the number of audio observations that are associated with a video frame, we consider a longer audio frame while we allow a large overlap between audio frames: this yields 30 feature vectors for each video frame.

A complex-valued feature vector is thus built from each window, whose module and argument describing the ILD (interaural level difference) and IPD (interaural phase difference) respectively. It is well known that these binaural cues contain sound direction information. Each feature vector is then mapped onto the image plane using the piecewise-affine high-dimensional to low-dimensional regression method of [Deleforge 15b]. In combination with voice activity detection (VAD), this process provides a time series of realizations of both the sound direction variables $\mathbf{Y}_{1:t}$ and the associated *speech-activity binary masks* $\mathbf{A}_{1:t}$, as detailed in Section 6.2.

In addition to our own data, we also tested our method on the dataset used in [Minotto 15]. These recordings contain one to three *static* persons *facing* the camera and the microphones, i.e., a Kinect. It is important to note that this dataset often contains persons that speak simultaneously and that speaker diarization is quite challenging in this case. Within this dataset, the *Two10* sequence is a representative example and hence we applied our method to this sequence. The audio recordings in this dataset used a microphone configuration quite different than ours, namely a linear microphone array with 8 microphones. For this reason we applied the SRP-PHAT sound-source localization method to the audio data available with the *Two10* sequence, which only provides the sound's azimuth; this direction is then mapped onto an image column using the microphone-to-camera transformation parameters of the Kinect, hence there is a large vertical sound-direction uncertainty.

We compared the proposed method with [Gebru 15a] and with [Minotto 15]. The main difference between the current work and [Gebru 15a] is the audio-visual association model. In [Gebru 15a] a GMM with a uniform component (GMM+U) is used while here we propose to use the weighted-data GMM (WD-GMM). Moreover, [Gebru 15a] considers a single audio observation for each video frame and the parameters of the GMM+U mixture are manually defined. The parameters of the proposed WD-GMM observation model are learned on-line by gathering audio observations within a 0.4 s window centered on each video frame. This robustly clusters audio observations generated by the same person. The diarization method proposed in [Minotto 15] uses a supervised classifier (SVM), trained using sequences from the same dataset (same acoustic environment), to discriminate between speaking and non-speaking persons. This contrasts with our online joint audio-visual observation model which is completely unsupervised.

The results in Table 6.1 reports speaker diarization scores in terms of Correct detection rates (CDR). The proposed model outperforms the one proposed in [Gebru 15a] for *counting* sequence, while it competes the state-of-the-art method of [Minotto 15], although the latter benefits from training on data from the same experimental setting.

Figures 6.1, 6.2 and 6.3 display our diarization results on the *counting*, *chat* and *Two10*

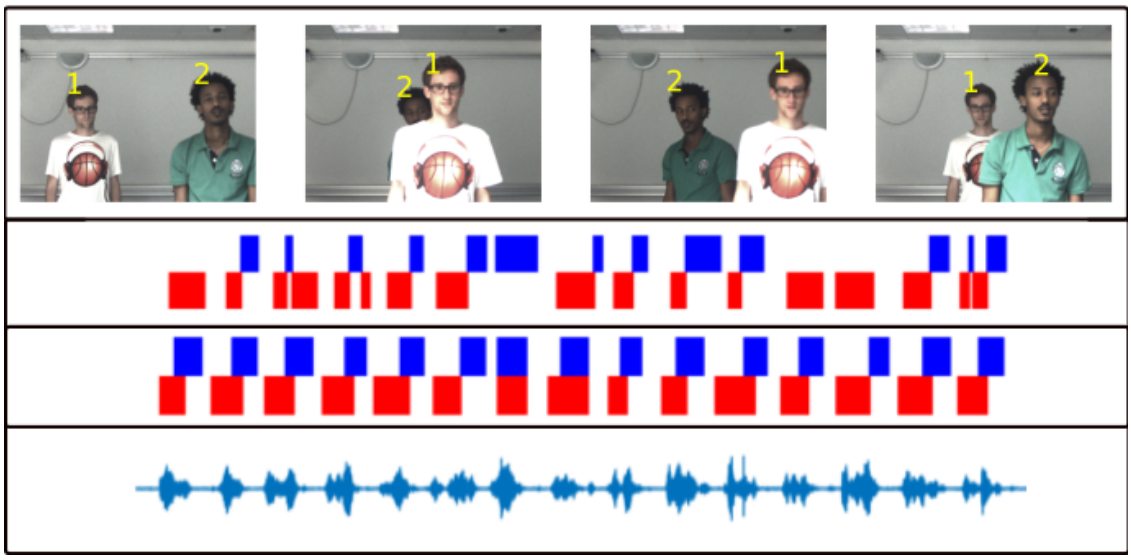


Figure 6.1: The *counting* sequence involves two moving persons that occasionally occlude each other. Visual tracking results (first row). Diarization results (second row) illustrated with a color diagram: each color corresponds to the audio activity of a person.Ground-truth diarization (third row); notice that there is a systematic overlap between the two speech signals. The raw audio signal delivered by the left microphone (fourth row).

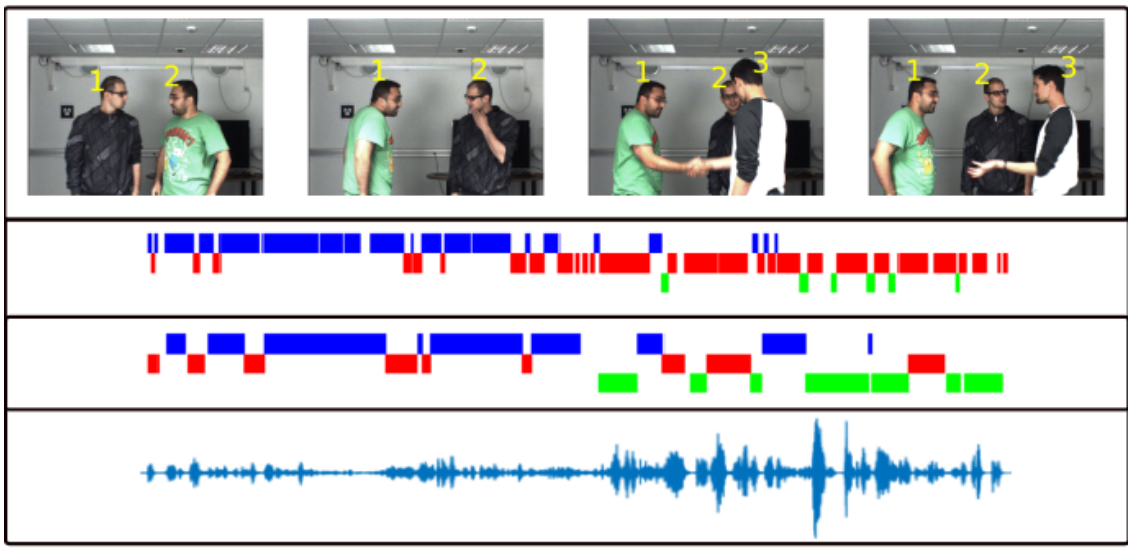


Figure 6.2: The *chat* sequence involves two then three moving persons that take speech turns and that occasionally occlude each other.

sequences, respectively. The proposed method obtains very good results over the *counting* sequence (see Figure 6.1) even if the sequence exhibits large portions where the two speakers speak at the same time. The performance over the challenging case of the *chat* sequence is lower than for the *counting* sequence. This drop can be explained by the fact

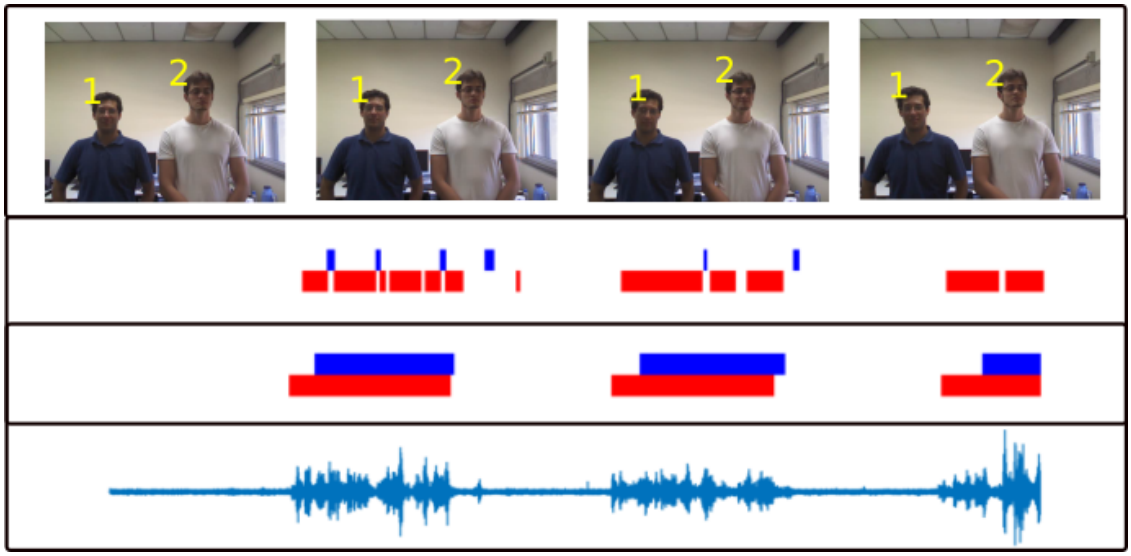


Figure 6.3: The *Two10* sequence from [Minotto 15] involves two static persons that speak simultaneously and always face the camera and the microphones.

that one speaker is mostly facing away both the camera and the microphones, thus his localization from audio data is much more challenging because of reverberations. Finally, the results on sequence *Two10* (Fig. 6.3) should be interpreted on the premise that our method detects only one speaker at a time.

Table 6.1: Correct detection rates (CDR) obtained by the proposed method and two other methods. The *Chat* and *Two10* sequences contain overlapping speaking persons. The *Chat* sequence contains a varying number of persons that take speech turns.

Sequence	Proposed	[Gebru 15a]	[Minotto 15]
Counting (Fig. 6.1)	84%	75%	n/a
Chat (Fig. 6.2)	55%	64%	n/a
Two10 (Fig. 6.3)	88%	n/a	92%

6.4 CONCLUSIONS

This chapter addressed the problem of active speaker tracking using auditory and visual data gathered with two microphones and one camera. Recent work in audio-visual diarization has capitalized on temporal coincidence of the two modalities, e.g., [Anguera Miro 12, Noulas 12]. In contrast, we propose active-speaker tracking method that enforces spatial coincidence: it exploits that a sound-source and associated visual-object should have the same spatial location. Consequently, it is possible to perform speaker localization by detecting and localizing persons in an image, estimating the directions of arrival of the active sound sources, mapping these sound directions onto the image, and associating the dominant sound source with one of the persons that are visible in the image. Moreover, this process is plugged into a dynamic Bayesian framework that robustly tracks the iden-

tity of the speakers and estimates a speech-turn latent variable. We described in detail the proposed method and illustrated its effectiveness with challenging scenarios involving moving people who speak inside a reverberant room and who may visually occlude each other.

CHAPTER 7

SPEAKER DIARIZATION BASED ON AV SPATIOTEMPORAL FUSION

Speaker diarization consists of assigning speech signals to people engaged in a dialogue. An audio-visual spatiotemporal diarization model is proposed. The model is well suited for challenging scenarios that consist of several participants engaged in multi-party interaction while they move around and turn their heads towards the other participants rather than facing the cameras and the microphones. Multiple-person visual tracking is combined with multiple speech-source localization in order to tackle the speech-to-person association problem. The latter is solved within a novel audio-visual fusion method on the following grounds: binaural spectral features are first extracted from a microphone pair, then a supervised audio-visual alignment technique maps these features onto an image, and finally a semi-supervised clustering method assigns binaural spectral features to visible persons. Moreover, the proposed model can process speech signals uttered simultaneously by multiple persons in a principled way. The diarization itself is cast into a latent-variable temporal graphical model that infers speaker identities and speech turns, based on the output of an audio-visual association process executed at each time step, and on the dynamics of the diarization variable itself. The proposed formulation yields an efficient exact inference procedure. We report extensive set of experiments and comparisons with respect to several state-of-the art diarization algorithms.

7.1 INTRODUCTION

In human-computer interaction (HCI) and human-robot interaction (HRI) it is often necessary to solve multi-party dialogue problems. For example, if two or more persons are engaged in a conversation, one important task to be solved, prior to automatic speech recognition (ASR) and natural language processing (NLP), is to correctly assign temporal segments of speech to corresponding speakers. In the speech and language processing literature this problem is referred to as *speaker diarization*, or "*who speaks when*?" A

number of diarization methods were recently proposed, *e.g.,* [Anguera Miro 12]. If only unimodal data are available, the task is extremely difficult. Acoustic data are inherently ambiguous because they contain mixed speech signals emitted by several persons, corrupted by reverberations, by other sound sources and by background noise. Likewise, the detection of speakers from visual data is very challenging and it is limited to lip and facial motion detection from frontal close-range images of people: in more general settings, such as informal gatherings, people are not always facing the cameras, hence lip reading cannot be readily achieved.

Therefore, an interesting and promising alternative consists of combining the merits of audio and visual data. The two modalities provide complementary information and hence audio-visual approaches to speaker diarization are likely to be more robust than audio-only or vision-only approaches. Several audio-visual diarization methods have been investigated for the last few years, *e.g.,* [Garau 10, Noulas 12, El Khoury 14, Minotto 15, Sarafianos 16, Kapsouras 16]. Diarization is based on audio-visual association, on the premise that a speech signal *coincides* with the visible face of a speaker. This coincidence must occur both in space and time.

In formal scenarios, *e.g.,* meetings, diarization is facilitated by the fact that participants take speech turns, which results in (i) a clear-cut distinction between speech and non-speech and (ii) the presence of short silent intervals between speech segments. Moreover, participants are seated, or are static, and there are often dedicated close-field microphones and cameras for each participant *e.g.,* [Carletta 05]. In these cases, the task consists of associating audio signals that contain clean speech with frontal images of faces: audio-visual association methods based on *temporal coincidence* between the audio and visual streams seem to provide satisfactory results, *e.g.,* canonical correlation analysis (CCA) [Kidron 05, Kidron 07, Sargin 07] or mutual information (MI) [Hershey 00, Fisher III 00, Garau 10, Noulas 12]. Nevertheless, temporal association between the two modalities is only effective on the premises that (i) speech segments are uttered by a single person at a time, that (ii) single-speaker segments are relatively long, and that (iii) speakers continuously face the cameras.

Moreover, in informal scenarios, *e.g.,* ad-hoc social events, the audio signals are provided by distant microphones, hence the signals are corrupted by environmental noise and by reverberations. Speakers interrupt each other, hence short speech signals may occasionally be uttered simultaneously by different speakers. Moreover, people often wander around, turn their head away from the cameras, may be occluded by other people, suddenly appear or disappear from the cameras' fields of view, etc. Some of these problems were addressed in the framework of audio-visual speaker tracking [Gatica-Perez 07, Naqvi 10a, Kilic 15a]. Nevertheless, audio-visual tracking is mainly concerned with finding speaker locations and speaker trajectories, rather than solving the speaker diarization problem.

In this chapter, we proposed a novel spatiotemporal diarization model that is well suited for challenging scenarios that consist of several participants engaged in multi-party dialogue. The participants are allowed to move around and to turn their heads towards the other participants rather than facing the cameras. We propose to combine multiple-person

visual tracking with multiple speech source localization in order to tackle the speech to person association problem. The latter is solved within a novel audio-visual fusion method on the following grounds: acoustic spectral features are extracted from a microphone pair, a novel supervised audio-visual alignment technique maps these features onto the image plane such that the audio and visual modalities are represented in the same mathematical space, a semi-supervised clustering method assigns the acoustic features to visible persons. The main advantage of this method over previous works (see Chapter 6, [Gebru 15a] and [Gebru 15b]) is twofold: it processes in a principled way speech signals uttered simultaneously by multiple persons, and it enforces spatial coincidence between audio and visual features.

Moreover, we cast the diarization process into a latent-variable temporal graphical model that infers over time both speaker identities and speech turns. This inference is based on combining the output of the proposed audio-visual fusion, that occurs at each time-step, with a dynamic model of the diarization variable (from the previous time-step to the current time-step), *i.e.,* a state transition model. We describe in detail the proposed formulation which is efficiently solved via an exact inference procedure. We tested the proposed model on a number of scenarios from the **AVDIAR** dataset (see Chapter 2 and [Gebru 17a]) involving several participants engaged in formal and informal dialogue. We also throughly benchmark the proposed method with respect to several state-of-the art diarization algorithms.

The remainder of this chapter is organized as follows. Section 7.2 describes the related work. Section 7.3 describes in detail the temporal graphical model. Section 7.4 describes visual feature detection and Section 7.5 describes the proposed audio features and their detection. Section 7.6 describes the proposed semi-supervised audio-visual association method. Numerous experiments and benchmarks are presented in Section 7.7. Finally, Section 7.8 draws some conclusions. Videos, Matlab code and additional examples are available online.[1]

7.2 Related Work

The task of speaker diarization is to detect speech segments and to group segments that correspond to the same speaker without any prior knowledge about the speakers involved nor their number. This can be done using auditory features alone, or a combination of auditory and visual features. Mel frequency cepstral coefficients (MFCC) is often the representation of choice whenever audio signal segments correspond to a single speaker. The diarization pipeline consists of splitting the audio frames into speech and non-speech frames, of extracting an MFCC feature vector from each speech frame and of performing agglomerative clustering such that each cluster found at the end corresponds to a different speaker [Wooters 08]. Consecutive speech frames are assigned either to the same speaker and grouped into segments, or to different speakers, by using a state transition model, *e.g.,* HMM.

The use of visual features for diarization has been motivated by the importance of

[1] `https://team.inria.fr/perception/avdiarization/`

audio-visual synchrony. Indeed, it was shown that facial and lip movements are strongly correlated with speech production [Yehia 98] and hence visual features, extracted from frontal views of speaker faces, can be used to increase the discriminative power of audio features in numerous tasks, *e.g.,* speech recognition [Potamianos 03], source separation, [Rivet 07, Barzelay 10] and diarization [Fisher III 00, Nock 03, Siracusa 07, Noulas 07]. In the latter case, the most common approaches involve the analysis of temporal correlation between the two modalities such that the face/lip movements that best correlate with speech correspond to an active speaker.

Garau et al. [Garau 10] compare two audio-visual synchronization methods that are based on MI and on CCA, and using MFCC auditory features combined with motion amplitude computed from facial feature tracks. They conclude that MI performs slightly better than CCA and that vertical facial displacements (lip and chin movements) are the visual features the most correlated with speech production. MI that combines gray-scale pixel-value variations extracted from a face region with acoustic energy is also used by Noulas et al. [Noulas 12]. The audio-visual features thus extracted are plugged into a dynamic Bayesian network (DBN) that perform speaker diarization. The method was tested on video meetings involving up to four participants which are recorded with several cameras, such that each camera faces a participant. More recently, both El Khoury et al. [El Khoury 14] and Kapsouras et al. [Kapsouras 16] propose to cluster audio features and face features independently and then to correlated these features based on temporal alignments between speech and face segments.

The methods mentioned so far yield good results whenever clean speech signals and frontal views of faces are available. A speech signal is said to be *clean* if it is noise free and if it corresponds to a single speaker; hence audio clustering based on MFCC features performs well. Moreover, time series of MFCC features seem to correlate well with facial-feature trajectories. If several faces are present, it is possible to select the facial-feature trajectory that correlate the most with the speech signal, e.g. [Kidron 05, Kidron 07]. However, in realistic settings, participants are not always facing the camera, consequently the detection of facial and lip movements is problematic. Moreover, methods based on cross-modal temporal correlation, e.g. [Fisher III 00, Nock 03, Potamianos 03, Siracusa 07, Noulas 07, Noulas 12] require long sequences of audiovisual data, hence they can only be used offline such as the analysis of broadcast news, of audiovisual conferences, etc.

In the presence of simultaneous speakers, the task of diarization is more challenging because multiple-speaker information must be extracted from the audio data, one one hand, and the speech-to-face association problem must be properly addressed, on the other hand. In mixed-speech microphone signals, or dirty speech, there are many audio frames that contain acoustic features uttered by several speakers and MFCC features are not reliable anymore because they are designed to characterize acoustic signals uttered by single speakers. The multi-speech-to-multi-face association problem cannot be solved neither by performing temporal correlation between a single microphone signal and an image sequence nor by clustering MFCC features.

One way to overcome the problems just mentioned is to perform multiple speech-source localization [Mandel 10, Blandin 12, Dorfan 15] and to associate speech sources

with persons. These methods, however, do not address the problems of aligning speech-source locations with visible persons and of tracking them over time. Moreover, they often use circular or linear microphone arrays, e.g. *planar* microphone setups, hence they provide sound-source directions with one degree of freedom, e.g. azimuth, which may not be sufficient to achieve robust audio-visual association. Hence, some form of microphone-camera calibration is needed. Khalidov et al. [Khalidov 11a] propose to estimate the microphone locations into a camera-centered coordinate system and to use a binocular-binaural setup in order to jointly cluster visual and auditory feature via a conjugate mixture model. Minotto et al. [Minotto 15] learn an SVM classifier using labeled audio-visual features. This training is dependent on the acoustic properties of experimental setup. They combine voice activity detection with sound-source localization using a linear microphone array which provides horizontal (azimuth) speech directions. In terms of visual features, their method relies on lip movements, hence frontal speaker views are required.

Multiple-speaker scenarios were thoroughly addressed in the framework of audio-visual tracking. Gatica-Perez et al. [Gatica-Perez 07] proposed a multi-speaker tracker using approximate inference implemented with a Markov chain Monte Carlo particle filter (MCMC-PF). Navqi et al. [Naqvi 10a] proposed a 3D visual tracker, based as well on MCMC-PF, to estimate the positions and velocities of the participants which are then passed to blind source separation based on beamforming [Van Veen 88].

In this chapter, we present a novel DBN-based multimodal speaker diarization model. Unlike several recently proposed multimodal diarization works [Noulas 12, El Khoury 14, Gebru 15a, Gebru 15b, Kapsouras 16], the proposed model can deal with simultaneously speaking participants that may wander around and turn their faces away from the cameras. Unlike [Noulas 12, El Khoury 14, Kapsouras 16] which require long sequences of past, present, and future frames, and hence are well suited for post-processing, our method is *causal* and therefore it can be used online. To deal with mixed speech signals, we exploit the sparsity of speech spectra and we propose a novel multiple speech-source localization method based on audio-visual data association implemented with a cohort of frequency-wise semi-supervised complex-Gaussian mixture models.

7.3 PROPOSED MODEL

We start by introducing a few notations and definitions. Unless otherwise specified, upper-case letters denote random variables while lower-case letters denote their realizations. Vectors are in slanted bold, *e.g.,* $\boldsymbol{X}, \boldsymbol{Y}$, while matrices are in bold, *e.g.,* $\mathbf{X}, \mathbf{Y}$. We consider an image sequence that is synchronized with two microphone signals and let t denote the time-step index of the audio-visual stream of data.

Let N be the maximum number of visual objects, *e.g.,* persons, available at any time t. Hence at t we have at most N persons with locations on the image plane $\mathbf{X}_t = (\boldsymbol{X}_{t,1}, \ldots, \boldsymbol{X}_{t,n}, \ldots, \boldsymbol{X}_{t,N}) \in \mathbb{R}^{2\times N}$, where the observed random variable $\boldsymbol{X}_{t,n} \in \mathbb{R}^2$ is the pixel location of person n at t. We also introduce a set of binary (or control) variables $\boldsymbol{V}_t = (V_{t,1}, \ldots V_{t,n}, \ldots V_{t,N}) \in \{0,1\}^N$ such that $V_{t,n} = 1$ if person n is *visible* at t

and $V_{t,n} = 0$ if the person is not visible. Let $N_t = \sum_n V_{t,n}$ denote the number of visible persons at t. The time series $\mathbf{X}_{1:t} = \{\mathbf{X}_1, \ldots, \mathbf{X}_t\}$ and associated *visibility binary masks* $\mathbf{V}_{1:t} = \{\boldsymbol{V}_1, \ldots, \boldsymbol{V}_t\}$ can be estimated using a multi-person tracker, *i.e.,* Section 7.4.

We now describe the audio data. Without loss of generality, the audio signals are recorded with two microphones: let $\mathbf{Y}_t = (\boldsymbol{Y}_{t,1}, \ldots, \boldsymbol{Y}_{t,k}, \ldots, \boldsymbol{Y}_{t,K}) \in \mathbb{C}^{F \times K}$ be a *binaural spectrogram* containing F number of frequencies and K number of frames. Each frame is a binaural vector $\boldsymbol{Y}_{t,k} \in \mathbb{C}^F, 1 \leq k \leq K$. Binaural spectrograms are obtained in the following way. The short-time Fourier transform (STFT) is first applied to the left- and right-microphone signals acquired at time-step t such that two spectrograms, $\mathbf{L}_t, \mathbf{R}_t \in \mathbb{C}^{F \times K}$ are associated with the left and right microphones, respectively. Each spectrogram is composed of $F \times K$ complex-valued STFT coefficients. The binaural spectrograms $\mathbf{Y}_t$ is composed of $F \times K$ complex-valued coefficients and each coefficients $Y_{t,k}^f, 1 \leq f \leq F$ and $1 \leq k \leq K$, can be estimated from the corresponding left- and right-microphone STFT coefficients $L_{t,k}^f$ and $R_{t,k}^f$, *i.e.,* Section 7.5. One important characteristic of speech signals is that they have sparse spectrograms. As explained below, this sparsity is explicitly exploited by the proposed speech-source localization method. Moreover, the microphone signals are obviously contaminated by background noise and by sounds emitted by other non-speech sources. Therefore, *speech activity* associated with each binaural spectrogram entry $Y_{t,k}^f$ must be properly detected and characterized with the help of a binary-mask matrix $\mathbf{A}_t \in \{0,1\}^{F \times K}$: $A_{t,k}^f = 1$ if the corresponding spectrogram coefficient contains speech, and $A_{t,k}^f = 0$ if it does not contain speech. To summarize, the binaural spectrograms $\mathbf{Y}_{1:t} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_t\}$ and associated *speech-activity masks* $\mathbf{A}_{1:t} = \{\mathbf{A}_1, \ldots, \mathbf{A}_t\}$ characterize the audio observations.

7.3.1 SPEAKER DIARIZATION MODEL

We remind that the objective of our work is to assign speech signal to persons, which amounts to one-to-one spatiotemporal associations between several speech sources (if any) and one or several observed persons. For this purpose we introduce a time series of discrete latent variables, $\mathbf{S}_{1:t} = \{\boldsymbol{S}_1, \ldots, \boldsymbol{S}_t\} \in \{0,1\}^{N \times t}$ where the vector $\boldsymbol{S}_t = (S_{t,1}, \ldots, S_{t,n}, \ldots, S_{t,N}) \in \{0,1\}^N$ has binary-valued entries such that $S_{t,n} = 1$ if person n *speaks* during the time-step t, and $S_{t,n} = 0$ if person n is *silent*. The temporal speaker diarization problem at hand can be formulated as finding a maximum-a-posteriori (MAP) solution, namely finding the most probable configuration of the latent state $\boldsymbol{S}_t$ that maximizes the following posterior probability distribution, also referred to as the filtering distribution:

$$\hat{\boldsymbol{s}}_t = \underset{\boldsymbol{s}_t}{\operatorname{argmax}} \, \mathrm{P}(\boldsymbol{S}_t = \boldsymbol{s}_t | \mathbf{x}_{1:t}, \mathbf{y}_{1:t}, \mathbf{v}_{1:t}, \mathbf{a}_{1:t}). \tag{7.1}$$

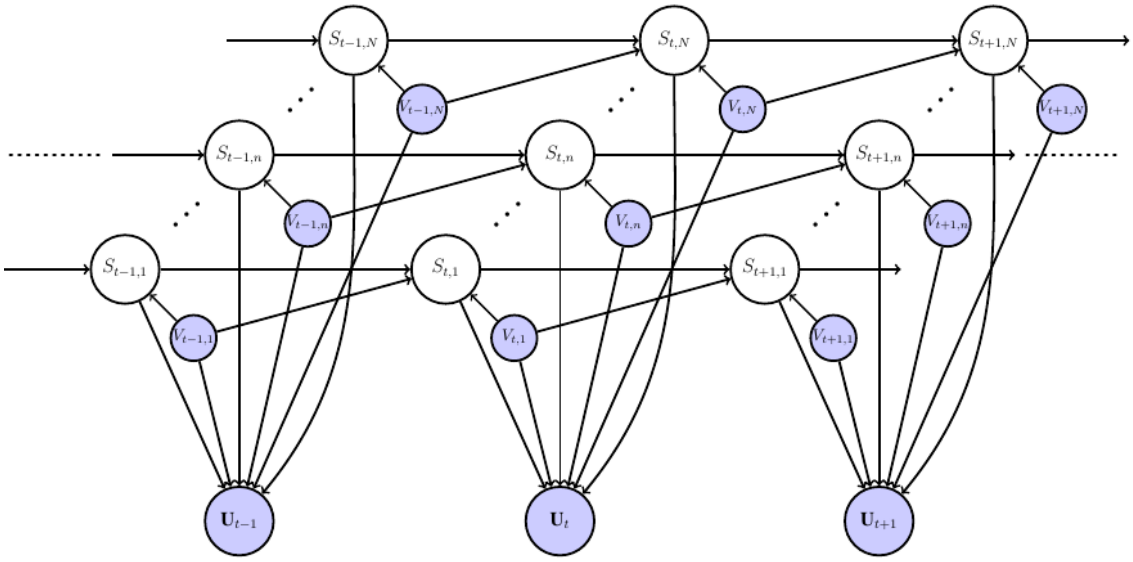


Figure 7.1: The Bayesian spatiotemporal fusion model for audio-visual speaker diarization. Shaded nodes represent the observed variables, while unshaded nodes represent latent variables. Note that the visibility-mask variables $V_{t,n}$ although observed, they are treated as control variables. This model enables simultaneously speaking persons, which is not only a realistic assumption but also very common in natural dialogues and applications like for example HRI.

We introduce the notation $\mathbf{U}_t = (\mathbf{X}_t, \mathbf{Y}_t, \mathbf{A}_t)$ for the observed variables, while the $\mathbf{V}_t$ are referred to as control variables. The filtering distribution (7.1) can be expanded as:

$$\begin{aligned} \mathrm{P}(\mathbf{s}_t|\mathbf{u}_{1:t}, \mathbf{v}_{1:t}) &= \frac{\mathrm{P}(\mathbf{u}_t|\mathbf{s}_t, \mathbf{u}_{1:t-1}, \mathbf{v}_{1:t})\mathrm{P}(\mathbf{s}_t|\mathbf{u}_{1:t-1}, \mathbf{v}_{1:t})}{\mathrm{P}(\mathbf{u}_t|\mathbf{u}_{1:t-1}, \mathbf{v}_{1:t})} \\ &= \frac{\mathrm{P}(\mathbf{u}_t|\mathbf{s}_t, \boldsymbol{v}_t)\mathrm{P}(\mathbf{s}_t|\mathbf{u}_{1:t-1}, \mathbf{v}_{1:t})}{\mathrm{P}(\mathbf{u}_t|\mathbf{u}_{1:t-1}, \mathbf{v}_{1:t})}. \end{aligned} \tag{7.2}$$

We assumed that the observed variables $\mathbf{U}_t$ are conditionally independent of all other variables, given the speaking state $\boldsymbol{S}_t$ and control input $\mathbf{V}_t$; $\boldsymbol{S}_t$ is conditionally independent of $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_{t-2}$, given $\boldsymbol{S}_{t-1}$ and $\boldsymbol{V}_{t-1:t}$. Fig. 7.1 shows the graphical model representation of the proposed model.

The numerator of (7.2) is the product of two terms: the observation likelihood (left) and the predictive distribution (right). The observation likelihood can be expanded as:

$$\mathrm{P}(\mathbf{u}_t|\mathbf{s}_t, \boldsymbol{v}_t) = \prod_{n=1}^{N} \mathrm{P}(\mathbf{u}_t|S_{t,n} = 1, V_{t,n})^{s_{t,n}} \mathrm{P}(\mathbf{u}_t|S_{t,n} = 0, V_{t,n})^{1-s_{t,n}}. \tag{7.3}$$

The predictive distribution (right hand side of the numerator of (7.2) expands as:

$$\mathrm{P}(\mathbf{s}_t|\mathbf{u}_{1:t-1},\mathbf{v}_{1:t}) = \sum_{\mathbf{s}_{t-1}} \mathrm{P}(\mathbf{s}_t,\mathbf{s}_{t-1}|\mathbf{u}_{1:t-1},\mathbf{v}_{1:t}) \tag{7.4}$$

$$= \sum_{\mathbf{s}_{t-1}} \mathrm{P}(\mathbf{s}_t|\mathbf{s}_{t-1},\mathbf{u}_{1:t-1},\mathbf{v}_{1:t})\mathrm{P}(\mathbf{s}_{t-1}|\mathbf{u}_{1:t-1},\mathbf{v}_{1:t}) \tag{7.5}$$

$$= \sum_{\mathbf{s}_{t-1}} \mathrm{P}(\mathbf{s}_t|\mathbf{s}_{t-1},\boldsymbol{v}_t,\boldsymbol{v}_{t-1})\mathrm{P}(\mathbf{s}_{t-1}|\mathbf{u}_{1:t-1},\mathbf{v}_{1:t-1}) \tag{7.6}$$

$$= \sum_{\mathbf{s}_{t-1}} \left(\prod_{m=1}^{N} \mathrm{P}(s_{t,m}|s_{t-1,m},v_{t,m},v_{t-1,m}) \right) \times \mathrm{P}(\mathbf{s}_{t-1}|\mathbf{u}_{1:t-1},\mathbf{v}_{1:t-1}). \tag{7.7}$$

(7.7) is the product of the state transition probabilities and of the filtering distribution at $t-1$.

We now expand the denominator of (7.2):

$$\mathrm{P}(\mathbf{u}_t|\mathbf{u}_{1:t-1},\mathbf{v}_{1:t}) = \sum_{\mathbf{s}_t} \mathrm{P}(\mathbf{u}_t,\mathbf{s}_t|\mathbf{u}_{1:t-1},\mathbf{v}_{1:t})$$

$$= \sum_{\mathbf{s}_t} \mathrm{P}(\mathbf{u}_t|\mathbf{s}_t,\mathbf{v}_t)\mathrm{P}(\mathbf{s}_t|\mathbf{u}_{1:t-1},\mathbf{v}_{1:t}). \tag{7.8}$$

To summarize, the evaluation of the filtering distribution at an arbitrary time-step t requires the evaluation of (i) the observation likelihood (7.3), *i.e.,* Section 7.6, (ii) the state transition probabilities (7.6), *i.e.,* Section 7.3.2, (iii) the filtering distribution at $t-1$ (7.7), and of (iv) the normalization term (7.8). Notice that the number of possible state configuration is 2^N where N is the maximum number of people. For small values of N (2 to 6 persons), solving the MAP problem (7.1) is computationally efficient.

7.3.2 State Transition Model

Priors over the dynamics of the state variables in (7.6) exploit the simplifying assumption that the speaking dynamics of a person is independent of all the other persons, *i.e.,*

$$\mathrm{P}(\mathbf{s}_t|\mathbf{s}_{t-1},\boldsymbol{v}_t,\boldsymbol{v}_{t-1}) = \prod_{n=1}^{N} \mathrm{P}(s_{t,n}|s_{t-1,n},v_{t,n},v_{t-1,n}). \tag{7.9}$$

Nevertheless, several existing speech-turn models rely on non-verbal cues, such as filled pauses, breath, facial gestures, gaze, etc. [Bohus 11, Skantze 14], and a speech-turn classifier can be built from annotated dialogues. In [Noulas 12] the state transition model considers all possible transitions, e.g., speaking/non-speaking, visible/not-visible, etc., which results in a large number of parameters that need be estimated. These models cannot be easily extended when there are speech overlaps and one has to rely on features extracted from the data.

To define the speaking transition priors $\mathrm{P}(s_{t,n}|s_{t-1,n}, v_{t,n}, v_{t-1,n})$, we consider three cases: (i) person n visible at $t-1$ and visible at t, or $v_{t,n} = v_{t-1,n} = 1$ and in this case the transitions are parametrized by a self-transition prior $q \in [0, 1]$ which models the probability to remain in the same state, either speaking or not speaking, (ii) person n not visible at $t-1$ and visible at t, or $v_{t,n} = 1, v_{t-1,n} = 0$, in this case, the prior to be either speaking or not speaking at t is uniform, and (iii) person n not visible at t, or $v_{t,n} = 0, v_{t-1,n} = 1$, in which case the prior not to be speaking is equal to 1. The following equation summarizes all these cases:

$$\mathrm{P}(s_{t,n}|s_{t-1,n}, v_{t,n}, v_{t-1,n}) = v_{t,n}v_{t-1,n}q^{\delta_{s_{t-1,n}}(s_{t,n})}(1-q)^{1-\delta_{s_{t-1,n}}(s_{t,n})} + \frac{1}{2}(1-v_{t-1,n})v_{t,n} + (1-v_{t,n})\delta_0(s_{t,n}), \tag{7.10}$$

where $\delta_i(j) = 1$ if $i = j$ and $\delta_i(j) = 0$ if $i \neq j$. Note that this does not consider the case of person n not visible at $t-1$ and at t for which the prior probability to be speaking is 0. In all our experiments we used $q = 0.8$.

The multiple-speaker tracking and diarization model proposed in this work only considers persons that are both seen and heard. Indeed, in informal scenarios there may be acoustic sources (speech or other sounds such as music) that are neither in the camera field of view, nor can they be visually detected and tracked. The proposed audio-visual association model addresses this problem, i.e. Section 7.6.

7.4 VISUAL OBSERVATIONS

We propose to use visual tracking of multiple persons in order to infer realizations of the random variables $\mathbf{X}_{1:t}$ introduced above. The advantage of a multiple-person tracker is that it is able to detect a variable number of persons, possibly appearing and disappearing from the visual field of view, to estimate their velocities, and to track their locations and identities. Multiple object/person tracking is an extremely well studied topic in the computer vision literature and many methods with their associated software packages are available. Among all these methods, we chose the multiple-person tracker of [Bae 14]. In the context of our work, this method has several advantages: (i) it robustly handles fragmented tracks (due to occlusions, to the limited camera field of view, or simply to unreliable detections), (ii) it handles changes in person appearance, such as a person that faces the camera and then suddenly turns his/her head away from the camera, *e.g.,* towards a speaker, and (iii) it performs online discriminative learning such that it can distinguish between similar appearances of different persons.

Visual tracking is implemented in the following way. Upper-body detector proposed in [Bourdev 09] is used to extract bounding boxes of persons in every frame. This allows the tracker to initialize new tracks, to re-initialize lost ones, to avoid tracking drift, and to cope with a large variety of poses and resolutions. Moreover, an appearance model, based on the color histogram of a bounding box associated with a person upper body (head and torso), is associated with each detected person. The appearance model is updated whenever the upper-body detector returns a reliable bounding box (no overlap with another bounding box). We observed that upper-body detection is more robust than face detection

which yields many false positives. Nevertheless, in the context of audio-visual fusion, the face locations are important. Therefore, the locations estimated by the tracker, $\mathbf{X}_{1:t}$, correspond to the face centers of the tracked persons.

7.5 AUDIO OBSERVATIONS

In this section we present a methodology for extracting binaural features in the presence of either a single audio source or several speech sources. We consider audio signals recorded with a binaural microphone pair. As already explained in Section 7.3, the short-time Fourier transform (STFT) is applied to the two microphone signals acquired at time-slice t and two spectrograms are thus obtained, namely $\mathbf{L}_t, \mathbf{R}_t \in \mathbb{C}^{F \times K}$.

7.5.1 SINGLE AUDIO SOURCE

Let's assume that there is a single (speech or non-speech) signal emitted by an audio source during the time slice t. In the STFT domain, the relationships between the source-STFT spectrogram and microphone-STFT spectrograms are, for each frame k and each frequency f (for convenience we omit the time index t):

$$L_k^f = H_{L,k}^f T_k^f + N_{L,k}^f \tag{7.11}$$

$$R_k^f = H_{R,k}^f T_k^f + N_{R,k}^f, \tag{7.12}$$

where $\mathbf{T} = \{T_k^f\}_{k=1,f=1}^{k=K,f=F}$ is the unknown source spectrogram, $\mathbf{N}_L = \{N_{L,k}^f\}_{k=1,f=1}^{k=K,f=F}$ and $\mathbf{N}_R = \{N_{R,k}^f\}_{k=1,f=1}^{k=K,f=F}$ are the unknown noise spectrograms associated with the left and right channels, and $\mathbf{H}_L = \{H_{L,k}^f\}_{k=1,f=1}^{k=K,f=F}$ and $\mathbf{H}_R = \{H_{R,k}^f\}_{k=1,f=1}^{k=K,f=F}$ are the unknown left and right *acoustic transfer functions* (ATF) that are frequency-dependent. The above equations correspond to the general case of a moving sound source. However, if we assume that the audio source is static during the time slice t, *i.e.,* the source emitter is in a fixed position during the time slice t, the acoustic transfer functions are time-invariant and only depend on the source position relative to the microphones. We further define *binaural features*, *i.e.,* the ratio between the left and right acoustic transfer functions, H_L^f / H_R^f. Notice that we omitted the frame index because in the case of a static source, the acoustic transfer function is invariant over frames. Likewise the acoustic transfer function, the binaural features do not depend on k and they only contain audio-source position information [Deleforge 15b].

One can use the estimated cross-PSD (power spectral density) and auto-PSD to extract binaural features in the following way. The cross-PSD between the two microphones is [Li 15a, Li 15b]:

$$\Phi_{L,R}^f = \frac{1}{K} \sum_{k=1}^{K} L_k^f R_k^{f\star} \tag{7.13}$$

$$\approx \frac{1}{K} H_L^f H_R^{f\star} \sum_{k=1}^{K} |T_k^f|^2 + \frac{1}{K} \sum_{k=1}^{K} N_{L,k}^f N_{R,k}^{f*}, \tag{7.14}$$

where $A^\star$ is the complex-conjugate of A and it is assumed that the signal-noise cross terms can be neglected. If the noise signals are spatially uncorrelated then the noise-noise cross terms can also be neglected. The binaural feature vector at t can be approximated with the ratio between the cross-PSD and auto-PSD functions, *i.e.,* the vector $\boldsymbol{Y}_t = (Y_t^1, \dots Y_t^f, \dots Y_t^F)^\top$ with entries:

$$Y_t^f = \frac{\Phi_{t,L,R}^f}{\Phi_{t,R,R}^f}. \tag{7.15}$$

7.5.2 Multiple Speech Sources

We now consider the case of P speakers ($P > 1$) that emit speech signals simultaneously (for convenience we omit again the time index t)

$$L_k^f = \sum_{p=1}^{P} H_{L,p}^f T_{p,k}^f + N_{L,k}^f \tag{7.16}$$

$$R_k^f = \sum_{p=1}^{P} H_{R,p}^f T_{p,k}^f + N_{R,k}^f, \tag{7.17}$$

where $H_{L,p}^f$ and $H_{R,p}^f$ are the acoustic transfer functions from the speech-source p to the left and right microphones, respectively. The STFT based estimate of the cross-PSD for each frequency-frame point (f, k) is

$$\Phi_{L,R,k}^f = L_k^f R_k^{f\star}. \tag{7.18}$$

In order to further characterize simultaneously emitting speech signals, we exploit the well-known fact that speech signals have sparse spectrograms in the Fourier domain. Because of this sparsity it is realistic to assume that only one speech source p is active at each frequency-frame point of the two microphone spectrograms (7.16) and (7.17). Therefore these spectrograms are composed of STFT coefficients that contain (i) either speech emitted by a single speaker, (ii) or noise. Using this assumption, the binaural spectrogram $\mathbf{Y}_t$ and associated binary mask matrix $\mathbf{A}_t$ can be estimated from the cross-PSD and auto-PSD in the following way. We start by estimating a binary mask for each frequency-frame point,

$$A_k^f = \begin{cases} 0 & \text{if } \max(\Phi_{L,L,k}^f, \Phi_{R,R,k}^f) < a \\ 1 & \text{otherwise,} \end{cases} \tag{7.19}$$

where a is an adaptive threshold whose value is estimated based on noise statistics [Li 16]. Then, we compute the binaural spectrogram coefficients for each frequency-frame point (f, k) at time-step t as:

$$Y_{t,k}^f = \begin{cases} \frac{\Phi_{t,L,R,k}^f}{\Phi_{t,R,R,k}^f} & \text{if} \quad A_{t,k}^f = 1 \\ 0 & \text{if} \quad A_{t,k}^f = 0. \end{cases} \tag{7.20}$$

It is important to stress that while these binaural coefficients are source independent, they are location dependent. This is to say that the binaural spectrogram only contains information about the location of the sound source and not about the content of the source. This crucial property allows one to use different types of sound sources for training a sound source localizer and for predicting the location of a speech source, as explained in the next section.

7.6 AUDIO-VISUAL FUSION

In this section we propose an audio-visual spatial alignment model that will allow us to evaluate the observation likelihood (7.3). The proposed audio-visual alignment is weakly supervised and hence it requires training data. We start by briefly describing the audio-visual training data. The training data contain pairs of audio recordings and their associated directions. Let $\widetilde{\mathbf{W}} = \{\widetilde{\boldsymbol{W}}_1, \ldots, \widetilde{\boldsymbol{W}}_m, \ldots \widetilde{\boldsymbol{W}}_M\} \in \mathbb{C}^{F \times M}$ be a training dataset containing M binaural vectors. Each binaural vector is extracted from its corresponding binaural audio recording using the method outlined above in Section 7.5.1, *i.e.,* $\widetilde{\boldsymbol{W}}_m = (\widetilde{W}_m^1, \ldots, \widetilde{W}_m^f, \ldots, \widetilde{W}_m^F)$ where each entry $\widetilde{W}_m^f$ is computed with (7.15).

Each audio sample in the training set consists of a recording of white-noise signal that is emitted by a loudspeaker placed at different locations, e.g. Fig. 2.2. The PSD of a white-noise signal is significant at each frequency thus: $|\widetilde{W}_m^f|^2 > a > 0, \forall m \in [1 \ldots M], \forall f \in [1 \ldots F]$. A visual marker placed onto the loudspeaker allows to associate its pixel location with each sound direction, as described in Section 2.6. Hence the M source directions correspond to an equal number of pixel locations $\widetilde{\mathbf{X}} = \{\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_m, \ldots \widetilde{\boldsymbol{X}}_M\} \in \mathbb{R}^{2 \times M}$. To summarize, the training data consist of M pairs of binaural features and associated pixel locations: $\{\widetilde{\boldsymbol{W}}_m, \widetilde{\boldsymbol{X}}_m\}_{m=1}^M$.

We now consider the two sets of visual and auditory observations during the time-step t:

$$\left.\begin{array}{l} \mathbf{X}_t = (\boldsymbol{X}_{t,1}, \ldots, \boldsymbol{X}_{t,n}, \ldots, \boldsymbol{X}_{t,N}) \in \mathbb{R}^{2 \times N}, \\ \boldsymbol{V}_t = (V_{t,1}, \ldots V_{t,n}, \ldots V_{t,N}) \in \{0,1\}^N \end{array}\right\} \text{ visual observations} \tag{7.21}$$

$$\left.\begin{array}{l} \mathbf{Y}_t = (\boldsymbol{Y}_{t,1}, \ldots, \boldsymbol{Y}_{t,k}, \ldots, \boldsymbol{Y}_{t,K}) \in \mathbb{C}^{F \times K}, \\ \mathbf{A}_t \in \{0,1\}^{F \times K} \end{array}\right\} \text{ auditory observations} \tag{7.22}$$

If person n, located at $\boldsymbol{X}_{t,n}$, is both visible and speaks at t: the binaural features associated with the emitted speech signal depend on the person's location only, hence they must be similar to the binaural features of the training source emitting from the same location. This can be simply written as a nearest neighbor regression over the training-set of audio-source locations:

$$\widetilde{\boldsymbol{X}}_n = \underset{m}{\operatorname{argmin}} \|\boldsymbol{X}_{t,n} - \widetilde{\boldsymbol{X}}_m\|^2 \tag{7.23}$$

and let $\widetilde{\boldsymbol{W}}_n \in \widetilde{\mathbf{W}}$ be the binaural feature vector associated with this location. Hence, the training pair $\{\widetilde{\boldsymbol{X}}_n, \widetilde{\boldsymbol{W}}_n\} \in \widetilde{\mathbf{X}} \times \widetilde{\mathbf{W}}$ can be associated with person n.

We choose to model that at any frequency $f \in [1 \dots F]$, the likelihood of and observed binaural feature $Y_{t,k}^f$ follows the following complex-Gaussian mixture model (for convenience, we omit the the time index t)

$$\mathrm{P}(Y_k^f|\mathbf{\Theta}^f) = \sum_{n=1}^{N} \pi_n^f \mathcal{N}_c(Y_k^f|\widetilde{W}_n^f, \sigma_n^f) + \pi_{N+1}^f \mathcal{N}_c(Y_k^f|0, \sigma_{N+1}^f), \tag{7.24}$$

where $\mathcal{N}_c(x|\mu,\sigma) = (\pi\sigma)^{-1}\exp(-|x-\mu|^2/\sigma), \quad x \in \mathbb{C}$ is the complex-normal distribution and $\mathbf{\Theta}^f$ is the set of *real-valued* model parameters, namely the priors $\{\pi_n^f\}_{n=1}^{N+1}$ with $\sum_{n=1}^{N+1}\pi_n^f = 1$, and the variances $\{\sigma_n^f\}_{n=1}^{N+1}$. This model states that the binaural feature Y_k^f is either generated by one of the N persons, located at $\widetilde{\boldsymbol{X}}_n, 1 \le n \le N$, hence it is an inlier generated by a complex-normal mixture model with means $\widetilde{W}_n^f, 1 \le n \le N$, or is emitted by an unknown sound source, hence it is an outlier generated by a zero-centered complex-normal distribution with a very large variance $\sigma_{N+1}^f \gg \sigma_n$.

The parameter set $\mathbf{\Theta}^f$ of (7.24) can be easily estimated via a simplified variant of the EM algorithm for Gaussian mixtures: the algorithm alternates between **E-step** that evaluates the posterior probabilities $r_{kn}^f = \mathrm{P}(z_k^f = n|Y_k^f)$, z_k^f is assignment variable, $z_k^f = n$ means Y_k^f is generated by component n:

$$r_{kn}^f = \begin{cases} \frac{1}{C}\pi_n^f \mathcal{N}_c(Y_k^f|\widetilde{W}_n^f, \sigma_n^f) & \text{if } 1 \le n \le N \\ \frac{1}{C}\pi_{N+1}^f \mathcal{N}_c(Y_k^f|0, \sigma_{N+1}^f) & \text{if } n = N+1, \end{cases} \tag{7.25}$$

$$\text{where } C = \sum_{i=1}^{N} \pi_i^f \mathcal{N}_c(Y_k^f|\widetilde{W}_i^f, \sigma_i^f) + \pi_{N+1}^f \mathcal{N}_c(Y_k^f|0, \sigma_{N+1}^f),$$

and **M-step** that estimates the variances and the priors:

$$\sigma_n^f = \frac{\sum_{k=1}^{K} A_k^f r_{kn}^f |Y_k^f - \widetilde{W}_n^f|^2}{\sum_{k=1}^{K} A_k^f r_{kn}^f} \quad \forall n, 1 \le n \le N \tag{7.26}$$

$$\pi_n^f = \frac{\sum_{k=1}^{K} A_k^f r_{kn}^f}{\sum_{k=1}^{K} A_k^f} \quad \forall n, 1 \le n \le N+1. \tag{7.27}$$

The algorithm can be easily initialized by setting all the priors equal to $\frac{1}{N+1}$ and by setting all the variances equal to a positive scalar σ. Because the component means are fixed, the algorithm converges in only a few iterations.

Based on these results one can evaluate (7.3), namely the speaking probability of person n located at $\boldsymbol{X}_n$: the probability that a visible person either speaks:

$$\mathrm{P}(\mathbf{U}_t|S_{t,n} = 1, V_{t,n} = 1) = \frac{\sum_{f=1}^{F}\sum_{k=1}^{K} A_{t,k}^f r_{t,kn}^f}{\sum_{f=1}^{F}\sum_{k=1}^{K} A_{t,k}^f}, \tag{7.28}$$

or is silent:

$$\mathrm{P}(\mathbf{U}_t|S_{t,n} = 0, V_{t,n} = 1) = 1 - \mathrm{P}(\mathbf{U}_t|S_{t,n} = 1, V_{t,n}). \tag{7.29}$$

7.7 EXPERIMENTAL EVALUATION

7.7.1 AUDIO-VISUAL DATASETS

In order to evaluate the proposed model, we used the **AVDIAR** dataset that was purposively gathered to encompass a wide number of multiple-speaker scenarios (see Chapter 2 and [Gebru 17a]). Amongst the various scenarios available in the dataset we used fourteen different sequences with varying complexity. They are summarized in Table 7.1.

In addition to the **AVDIAR** dataset, we used three publicly available datasets, *e.g.,* Fig. 7.2. They are briefly described as follows:

- The **MVAD** dataset described in [Minotto 15]. The visual data were recorded with a Microsoft Kinect sensor at 20 FPS,[2] and the audio signals were recorded with a linear array of omnidirectional microphones sampled at 44100 Hz. The recorded sequences are from 40 s to 60 s long and contain one to three participants that speak in Portuguese. The speech and silence segments are 4 s to 8 s long. Since the diarization method proposed in [Minotto 15] requires frontal faces, the participants are facing the camera and remain static through all the recordings.
- The **AVASM** dataset contains both training and test recordings used to benchmark the single and multiple speaker localization method described in [Deleforge 15b]. The recording setup is similar to the one described above, namely a binaural acoustic dummy head with two microphones plugged into its ears and a camera placed underneath the head. The images and the audio signals were captured at 25 FPS and 44100 Hz, respectively. The recorded sequences contain up to two participants that face the camera and speak simultaneously. In addition, the dataset has audio-visual alignment data collected in a similar fashion as the **AVDIAR** dataset.
- The **AV16P3** dataset is designed to benchmark audio-visual tracking of several moving speakers without taking conversational scenario and diarization task into account [Lathoud 04]. The sensor setup used for these recordings is composed of three cameras attached to the room ceiling, and two circular eight-microphone arrays. The recordings include mainly dynamic scenarios, comprising a single, as well as multiple moving speakers. In all the recordings there is a large overlap between the speaker-turns.

These datasets contain a large variety of recorded scenarios, aimed at a wide range of application. *e.g.,* formal and informal interaction in meetings and gatherings, human-computer interaction, etc. Some of the datasets were not purposively recorded to benchmark diarization. Nevertheless they are challenging because they contain a large amount of overlap between speakers, hence they are well suited to test the limits and failures of diarization methods. Unlike recordings of formal meetings, which are composed on long single-speech segments with almost no overlap between the participants, the above datasets contain the following challenging situations *e.g.,* Table 7.1:

[2]Note that our method doesn't use the depth image available with this sensor

- The participants do not always face the cameras, moreover, they turn their heads while they speak or listen;
- The participants, rather then being static, move around and hence the tasks of tracking and diarization must be finely intertwined;
- In informal meetings participants interrupt each other and hence not only that there is no silence between speech segments, but the speech segments overlap each other, and
- Participants take speech turns quite rapidly which results in short-length speech segments, which makes audio-visual temporal alignment quite challenging.

Table 7.1: List of sequences from the **AVDIAR** dataset used to evaluate the proposed diarization model.

Recordings	**Description**
Seq01-1P-S0M1, Seq04-1P-S0M1 , Seq22-1P-S0M1	A single person moving randomly and alternating between speech and silence.
Seq37-2P-S0M0, Seq43-2P-S0M0	Two static participants taking speech turns.
Seq38-2P-S1M0, Seq40-2P-S1M0, Seq44-2P-S2M0	Two static participants speaking almost simultaneously, *i.e.,* there are large speech overlaps.
Seq20-2P-S1M1, Seq21-2P-S2M1	Two participants, wandering in the room and engaged in a conversation, sometime speaking simultaneously.
Seq12-3P-S2M1, Seq27-3P-S2M1	Three participants engaged in an informal conversation. They are moving around and sometimes they speak simultaneously.
Seq13-4P-S1M1, Seq32-4P-S1M1	Three to four participants engaged in a conversation. Sometimes they speak simultaneously and there are many short speech turns.

7.7.2 Diarization Algorithms and Setup

We compared our method with four methods: [Vijayasenan 12], [Minotto 15], [Barzelay 10], and [Gebru 15b]. These methods are briefly explained below:

- Vijayasenan et al. [Vijayasenan 12] (*DiarTK*) use audio information only. *DiarTK* allows the user to incorporate a large number of audio features. In our experiments and comparisons we used the following features: mel-frequency cepstral coefficients (MFCC), frequency-domain linear prediction (FDLP), time difference of arrival (TDOA), and modulation spectrum (MS). Notice that TDOA features can only be used with static sound-sources, hence we did not use TDOA in the case of moving speakers.
- Minotto et al. [Minotto 15] learn an SVM classifier based on based on labeled audio-visual features. Sound-source localization provides horizontal sound directions which are combined with the output of a mouth tracker.
- Barzelay et al. [Barzelay 10] calculate audio-visual correlations based on extracting *onsets* from both modalities and on aligning these onsets.

Figure 7.2: Example frames from different datasets. The **MVAD** dataset (top) contains recordings of one to three persons that always face the camera. The **AVASM** (middle) was design to benchmark audio-visual sound-source localization with two simultaneously speaking persons or with a moving speaker. The **AV16P3** dataset (bottom) contains recordings of simultaneously moving and speaking persons.

The method consists of detecting faces and on tracking face landmarks, such that each landmark yields a trajectory. Onset signals are then extracting from each one of these trajectory as well as from the microphone signal. These onsets are used to compare each visual trajectory with the microphone signal, and the trajectories that best match the microphone signal correspond to the active speaker. We implemented this method based on [Barzelay 10] since there is no publicly available code. Extensive experiments with this method revealed that frontal views of speakers are needed. Therefore, we tested this methods with all the sequences from the **MVAD** and **AVASM** datasets and on the sequences from the **AVDIAR** dataset featuring frontal images of faces.

- Gebru et al. [Gebru 15b] track the active speaker, provided that participants take speech turns with no signal overlap. Therefore, whenever two persons speak simultaneously, this method extracts the *dominant* speaker.

Additionally, we used the following multiple sound-source localization methods:

- *GCC-PHAT* which is based on detecting the local maxima of the generalized cross-correlation function [Brandstein 97]. We used the implementation from the BSS Locate Toolbox [Blandin 12].
- *TREM* which considers a regular grid of source locations and selects the most probable locations based on maximum likelihood: we used the Matlab code provided by the authors, [Dorfan 15].

GCC-PHAT and TREM were used in conjunction with the proposed diarization method using the **AVDIAR** dataset as well as the **MVAD** and **AV3P16** datasets.

Unfortunately, we were not able to compare our method with the diarization methods of [Garau 10, Noulas 12] for two reasons: first, these methods require long speech segments (of the order of 10 s), and second the associated software packages are not publicly available, which would have facilitated the comparison task.

7.7.3 Diarization Performance Measure

To effectively benchmark our model with state-of-the art methods, we use the diarization error rate (DER) to quantitatively measure the performance: *smaller the **DER** value, better the performance*. DER is defined by the NIST-RT evaluation testbed,[3] and corresponds to the percentage of audio frames that are not correctly assigned to one or more speakers, or to none of them in case of a *silent* frame. DER consists of the composition of the following measurements:

- False-alarm error, when speech has been incorrectly detected;
- Miss error, when a person is speaking but the method fails to detect the speech activity, and
- Speaker-labeling error, when a person-to-speech association does not correspond to the ground truth.

To compute DER, the MD-EVAL software package of NIST-RT is used, setting the forgiveness collar to a video frame of *e.g.,* 40 ms for 25 FPS videos.

7.7.4 Results and Discussion

The results obtained with the **MVAD**, AVASM, **AV16P3** and **AVDIAR** datasets are summarized in Table 7.2, Table 7.3, Table 7.4 and Table 7.5, respectively.

Overall, it can be noticed that the method of [Barzelay 10] is the least performing method. As explained above this method is based on detecting signal onsets in the two modalities and on finding cross-modal correlations based on onset coincidence. Unfortunately, the visual onsets are unable to properly capture complex speech dynamics. The *DiarTK* method of [Vijayasenan 12] is the second least performing method. This is mainly due to the fact that this method is designed to rely on long speech segments with almost no overlap between consecutive segments. Whenever several speech signals overlap, it is very difficult to extract reliable information with MFCC features, since the latter are designed to characterize clean speech. *DiarTK* is based on clustering MFCC features using a Gaussian mixture model. Consider, for example, MFCC feature vectors of dimension 19, extracted from 20 ms-long audio frames, and a GMM with diagonal covariance matrices. If it is assumed that a minimum of 50 samples are needed to properly estimate

[3] `http://www.nist.gov/speech/tests/rt/2006-spring/`

Table 7.2: DER scores obtained with MVAD dataset (in %).

Sequence	**DiarTK** [Vijayasenan 12]	[Minotto 15]	[Barzelay 10]	[Gebru 15b]	Proposed with *TREM* [Dorfan 15]	Proposed with *GCC-PHAT* [Blandin 12]
One7	21.16	8.66	89.90	5.82	0.91	1.06
One8	20.07	7.11	98.10	4.92	1.02	1.81
One9	22.79	9.02	94.60	13.66	0.98	1.58
Two1	23.50	6.81	94.90	16.79	2.87	26.00
Two2	30.22	7.32	90.60	23.49	3.13	13.70
Two3	25.95	7.92	94.50	25.75	8.30	20.88
Two4	25.24	6.91	84.10	20.23	0.16	11.20
Two5	25.96	8.30	90.80	25.02	4.50	29.67
Two6	29.13	6.89	96.70	16.89	6.11	23.57
Two9	30.71	11.95	96.90	15.59	2.42	34.28
Two10	25.32	8.30	95.50	21.04	3.27	15.15
Two11	27.75	6.12	84.60	21.22	6.89	18.05
Two12	45.06	24.60	80.40	39.79	12.00	34.60
Two13	49.23	27.38	64.10	25.11	14.49	48.70
Two14	27.16	28.81	81.10	25.75	6.43	59.10
Three1	27.71	9.10	95.80	47.56	6.17	52.63
Three2	27.71	9.10	89.20	49.15	13.46	49.66
Three3	29.41	5.93	91.50	47.78	13.57	49.09
Three6	36.36	8.92	79.70	40.92	12.89	37.78
Three7	36.24	14.51	86.20	47.35	11.74	40.40
Average	29.33	11.18	89.96	26.69	6.57	28.45

Table 7.3: DER scores obtained with AVASM dataset (in %).

Sequence	[Barzelay 10]	[Gebru 15b]	Proposed with *TREM* [Dorfan 15]	Proposed with *GCC-PHAT* [Blandin 12]	Proposed
Moving-Speaker-01	95.04	6.26	21.84	17.24	6.26
Two-Speaker-01	70.20	24.11	34.41	44.42	2.96
Two-Speaker-02	80.30	26.98	32.52	47.30	7.33
Two-Speaker-03	74.20	35.26	46.77	47.77	13.78
Average	79.94	23.15	33.89	39.18	7.58

the GMM parameters, speech segments of at least $50\times19\times20$ ms, or 19 s, are needed. Therefore it is not surprising that *DiarTK* performs poorly on all these datasets.

Table 7.2 shows that the method of [Minotto 15] performs much better than *DiarTK*. This is not surprising, since the speech turns taken by the participants in the **MVAD** dataset are very brief. Minotto et al. [Minotto 15] use a combination of visual features extracted form frontal views of faces (lip movements) and audio features (speech-source directions) to train an SVM classifier. The method fails whenever the participants do not face the camera, *e.g.,* sequences *Two12*, Two13 and *Two14*, where participants purposely occlude their faces several times throughout the recordings. The method proposed in this paper in combination with *TREM* achieves the best results on almost all the tested scenarios. This is due to the fact that the audio-visual fusion method is capable of associating very short speech segments with one or several participants. However, the performance of our method, with either *TREM* or *GCC-PHAT*, drops down as the number of people increases. This is mainly due to the limited resolution of multiple sound-source localization algorithms (of the order of 10° horizontally) and thus, it makes it difficult to disambiguate

two nearby speaking/silent persons. Notice that tracking the identity of the participants is performed by visual tracking, which is a trivial task for most of these recordings, since participants are mostly static.

Table 7.3 shows the results obtained with the **AVASM** dataset. In these recordings the participants speak simultaneously, with the exception of the *Moving-Speaker-01* recording. We do not report results obtained with **DiarTK** since this method yields non-meaningful performance with this dataset. The proposed method performs reasonable well in the presence of simultaneously speaking persons.

Table 7.4 shows results obtained with the **AV16P3** dataset. As with the **AVASM** dataset we were unable to obtain meaningful results with the **DiarTK** method. As expected the proposed method has the same performance as [Gebru 15b] in the presence of a single active speaker, *e.g., seq11-1p-0100* and *seq15-1p-0111*. Nevertheless, the performance of [Gebru 15b] rapidly degrades in the presence of two and three persons speaking almost simultaneously. Notice that this dataset was recorded to benchmark audio-visual tracking, not diarization.

Table 7.4: DER scores obtained with AV16P3 dataset (in %).

Sequence	[Gebru 15b]	Proposed with *TREM* [Dorfan 15]	Proposed with *GCC-PHAT* [Blandin 12]
seq11-1p-0100	3.50	3.25	12.18
seq15-1p-0111	3.29	3.29	25.28
seq18-2p-0101	23.54	7.69	9.13
seq24-2p-0111	43.21	17.39	46.50
seq40-3p-1111	26.98	8.51	21.03
Average	20.04	8.02	22.82

Table 7.5 shows the results obtained with the **AVDIAR** dataset. The content of each scenario is briefly described in Table 2.1. The proposed method outperforms all other methods. It is also interesting to notice that our full method performs better than with either *TREM* or *GCC-PHAT*. This is due to the robust semi-supervised audio-visual association method proposed above. Fig. 7.3, Fig. 7.4, and Fig. 7.5 illustrate the audio-visual diarization results obtained by our method with three scenarios.[4]

7.8 CONCLUSIONS

In this chapter we proposed an audio-visual diarization method well suited for challenging scenarios consisting of participants in a multi-party interaction. We proposed to combine multiple-person visual tracking with multiple speech-source localization in a principled spatiotemporal Bayesian fusion model. Indeed, the diarization process was cast into a latent-variable dynamic graphical model. We described in detail the derivation of the proposed model and we showed that, in the presence of a limited number of speakers (of the

[4] Videos illustrating the performance of the proposed method using these scenarios are available at `https://team.inria.fr/perception/avdiarization/`.

Table 7.5: DER scores obtained with AVDIAR dataset (in %).

Sequence	*DiarTK* [Vijayasenan 12]	[Barzelay 10]	[Gebru 15b]	Proposed with *TREM* [Dorfan 15]	Proposed with *GCC-PHAT* [Blandin 12]	Proposed
Seq01-1P-S0M1	43.19	-	14.36	61.15	72.06	3.32
Seq04-1P-S0M1	32.62	-	14.21	71.34	68.84	9.44
Seq22-1P-S0M1	23.53	-	2.76	56.75	67.36	4.93
Seq37-2P-S0M0	12.95	34.70	1.67	41.02	45.90	2.15
Seq43-2P-S0M0	76.10	79.90	23.25	46.81	56.90	6.74
Seq38-2P-S1M0	47.31	59.20	43.01	47.89	47.38	16.07
Seq40-2P-S1M0	48.74	51.80	31.14	42.20	44.62	14.12
Seq20-2P-S1M1	43.58	-	51.78	58.82	59.38	35.46
Seq21-2P-S2M1	32.22	-	27.58	63.03	60.52	20.93
Seq44-2P-S2M0	54.47	-	44.98	55.69	51.0	5.46
Seq12-3P-S2M1	63.67	-	26.55	28.30	61.20	17.32
Seq27-3P-S2M1	46.05	-	20.84	47.40	68.79	18.72
Seq13-4P-S1M1	47.56	-	43.57	28.49	48.23	29.62
Seq32-4P-S1M1	41.51	-	43.26	33.36	71.98	30.20
Average	43.82	56.40	27.78	48.72	58.87	15.32

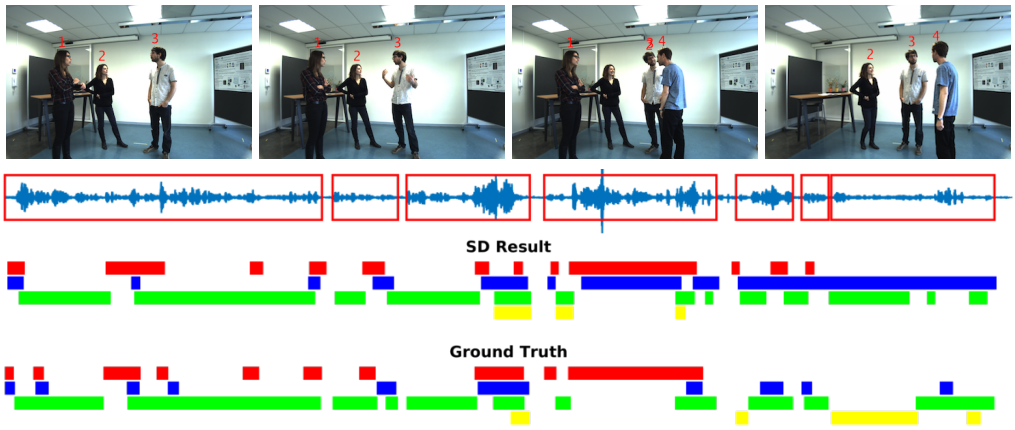


Figure 7.3: Results obtained on sequence Seq32-4P-S1M1. Visual tracking results (first row). The raw audio signal delivered by the left microphone and the speech activity region is marked with red rectangles (second row). Speaker diarization result (third row) illustrated with a color diagram: each color corresponds to the speaking activity of a different person. Annotated ground-truth diarization (fourth row).

order of ten), the diarization formulation is efficiently solved via an exact inference procedure. Then we described a novel multiple speech-source localization method and a weakly supervised audio-visual clustering method. We assess the performance of audio-visual (or audio-only) diarization methods using challenging scenarios, *e.g.,* the participants were allowed to freely move in a room and to turn their heads towards the other participants, rather than always facing the camera. We also benchmarked our method with several other recent methods using publicly available datasets.

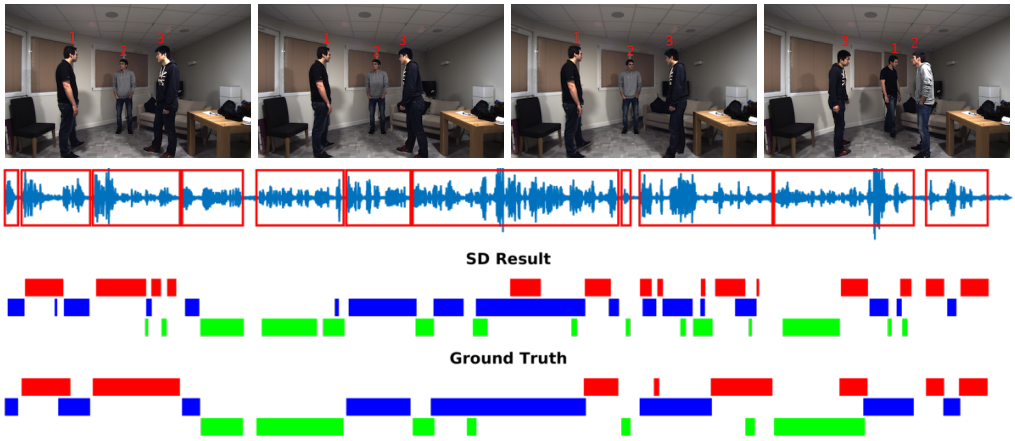


Figure 7.4: Result obtained on sequence Seq12-3P-S2M1.

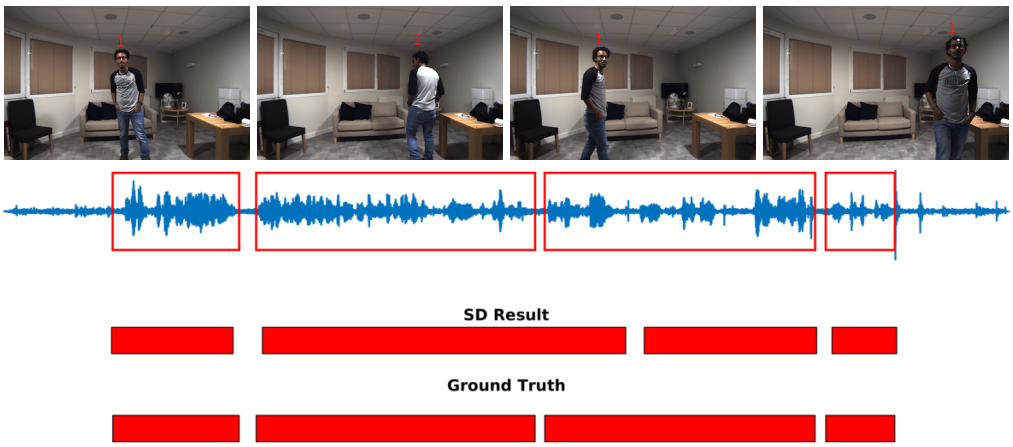


Figure 7.5: Result obtained on sequence Seq01-1P-S0M1.

This work can be extended in several ways. First, one could incorporate richer visual features, such as head pose estimation and head-pose tracking, in order to facilitate the detection of speech turns on the basis of gaze or of people that look at each other over time. One could also incorporate richer audio features, such as the possibility to extract speech signals emitted by each participant (sound-source separation) followed by speech recognition, and hence to enable not only diarization but also speech-content understanding. Another extension is to consider distributed sensors, wearable devices, or a combination of both, in order to be able to deal with more complex scenarios involving tens of participants [Yan 13, Alameda-Pineda 15b].

CHAPTER 8

CONCLUDING AND FUTURE DIRECTIONS

In this thesis we have focused on three tasks that stem from the challenging problem of audio-visual analysis for HRI: AV speaker localization, AV Multi-Person Tracking and AV Speaker Diarization. Our contributions aim at two main goals: (i) build robust models and algorithms for the robot that could provide audio-visual perception capabilities to achieve some level of natural interactivity with humans in natural and unconstrained environments; (ii) develop efficient techniques to combine/fuse/calibrate audio and video modalities in such a ways that one modality complement the weaknesses of the other modality. This chapter is organized as follows: Section 8.1 summarizes the contributions of the thesis; Section 8.2 gives directions for further research inspired from the work done in the thesis.

8.1 CONCLUDING REMARKS

Knowledge of the surrounding environment is prerequisite for interaction between humans and autonomous systems. In particular for human-robot-interaction (HRI), it is important for the robot to have the complete information of the humans around its vicinity, *e.g.,* their position, speaking status, body pose, gesture, emotion, facial expression, etc. Extracting any of these information by utilizing audio and video as separate cues is inefficient and the task becomes extremely difficult—since each modality necessarily has flaws or ambiguities. However, when both modalities are available, which is often the case since robots build for HRI are equipped with multiple acoustic and video sensors. Therefore, an interesting and promising alternative consists of combining the merits of audio and visual data. Within this context, we have investigated a number audio-visual research problems in this thesis, and below we would like to reempahsise the principal contributions.

In Chapter 2, we introduced the **AVDIAR** dataset which we have recored and made publicly available. The scenarios in the dataset cover examples of natural interaction between two or more people engaged in informal conversations. In an attempt to record

natural human-human interactions, all scenarios were unscripted and participants were allowed to enter, wander around, leave the scene at any time and interrupt each other while speaking. The acquisition setup, which consists of an acoustic dummy head, six microphones and two color cameras, was fully detailed. Technical specifications of the recorded streams (data) were provided. We annotated each video frame with bounding boxes of people faces and upper-body positions, as well as their identity and speaking activity over the entire sequence duration. We believe this dataset will provide an excellent test-bed for researchers working on multimodal speaker localization, tracking and diarization problems.

In Chapter 3 we have investigated the problem of weighted data clustering. We introduced weighted Gaussian mixture model (WD-GMM) and devised two EM algorithms to estimate model parameters. WD-GMM is useful to incorporate a prior weight information on the data instances. We proposed a data-driven weight initalization scheme if this weight information is not available through a prior expert knowledge. Furthermore, we proposed a model selection strategy based on a Minimum Message Length (MML) criterion. We have tested the proposed EM algorithms on several datasets and our experiments showed that the proposed EM algorithms compares favorably with several state-of-the-art parametric and non-parametric clustering methods and they perform particularly well in the presence of a large number of outliers.

In Chapter 4 we have investigated the problem of AV speaker localization. We proposed to address this in the framework of multimodal data clustering; using one of the EM algorithms proposed in Chapter 2. We described how audio-visual speaker localization problem may be cast into a challenging audio-visual data clustering problem, *e.g.,* how to associate human faces with speech signals and how to detect and localize active speakers in complex audio-visual scenes. The modules that translate raw audio and visual data into on-image observations are also described in detail. We then presented a robust technique that find the locations of speaking persons by exploiting on-image (spatial) coincidence of the visual and auditory observations. Moreover, we introduced a cross-modal weighting scheme in which observations from each modality are weighted accordingly to their relevance for speaker localization task. As a result of the cross-modal weighting scheme the proposed speaker localization technique is less affected by the presence of non-speaking persons, other sound sources, or in general noisy audio and visual observations.

In Chapter 5 we presented a novel approach to audio-visual tracking that jointly utilizes multiple DOAs obtained by sound source localization and facial detections to estimate the trajectories of multiple people in pixel space. The tracking problem is formulated in the framework of Bayesian filtering. Unlike classical approaches that use sampling techniques (*e.g.,* MCMC, PF) to represent the state-space posterior probability, we proposed to represent it with a mixture of Gaussian distributions (GMM). To ensure that a GMM representation can be retained sequentially over time, the predictive density is approximated by a GMM using the Unscented Transform (UT). Furthermore, a density interpolation technique is introduced to obtain a continuous representation of the audio-visual observations likelihood, which is also a GMM. The use of GMMs to represent the posterior, the predictive and the likelihood probability densities offer advantages of being able to model

multimodal distributions, handle nonlinear state transitions and to capture and model well the peculiarities of audio-visual observations. More importantly, a GMM can completely cover the whole state space, all with a modest number of parameters without requiring a large number of discrete samples. Furthermore, to prevent the number of mixtures from growing exponentially over time, a density approximation based on the WD-GMM EM algorithm proposed in Chapter 3 is applied, resulting in a compact GMM representation of the posterior density. Recordings using a camcorder and microphone array are used to evaluate the proposed approach, demonstrating significant improvements in tracking performance of the proposed audio-visual approach compared to two benchmark visual trackers.

In Chapter 6, we investigated the problem of speaker diarization. We cast the speaker diarization problem into a speaker tracking formulation whereby the active speaker is detected and tracked over time. We introduced a probabilistic model that exploits the on-image (spatial) coincidence of visual and auditory observations to robustly tracks the identity of the speakers and infers a single latent variable which represents the speaking activity (speech-turn). The spatial coincidence is build on the intuition that a sound-source and associated visual-object should have the same spatial location *i.e.,* spatial coincidence. Consequently, it is possible to perform speaker localization by detecting and localizing persons in an image, estimating the directions of arrival of the active sound sources, mapping these sound directions onto the image and associating the dominant sound source with one of the persons that are visible in the image. Hence, we plugged this process into the probabilistic model. We demonstrated the effectiveness of the proposed model with challenging scenarios involving moving people who speak inside a reverberant room and who may visually occlude each other.

In Chapter 7, we generalize the work presented in Chapter 6 to handle multiple simultaneous speakers. Consequently, we introduced a speaker diarization model based the fusion of spatio-temporal audio-visual data in Bayesian filtering framework. The model seamlessly combines auditory and visual data and is well suited for challenging scenarios that consist of several participants engaged in multi-party conversation, while they move around and turn their head towards the other participants rather than facing the cameras and the microphones. We combine multiple-person visual tracking is with multiple speech source localization in order to tackle the speech-to-person association problem. The latter is solved within a novel audio-visual fusion method on the following grounds: binaural spectral features are first extracted from a microphone pair, then a supervised audio-visual alignment technique maps these features onto images, and finally a semi-supervised clustering method assigns binaural spectral features to visible persons. The main advantage of this method is that it processes in a principled way speech signals uttered simultaneously by multiple persons. The diarization is cast into a latent-variable temporal graphical model that infers speaker identities and speech turns, based on the output of the audio-visual association process available at each discrete time-step, and on the dynamics of the diarization variable itself. The proposed formulation yields an efficient exact inference procedure, and is thoroughly tested and benchmarked with respect to several state-of-the art diarization algorithms.

8.2 DIRECTION FOR FUTURE RESEARCH

While thesis thesis focus on HRI settings using ego-centric microphones and camera setup, the algorithms and models developed can be easy ported to other settings that use network of distributed cameras and microphones *e.g.,* video-conferencing, surveillance, human activity understanding, etc. However, one has to develop appropriate audio-visual fusion strategies that best suits such configuration of audio-visual sensors. Furthermore, we would like to see the work done in this thesis as an opening to new research, we defined a non-exhaustive list of research directions to explore:

- This thesis mainly uses a single video camera and two microphones (also a spherical microphone array in Chapter 3. However, one may use more cameras and microphones. For example, the use of a stereo pair and depth sensors will allows the extraction of depth information. The use of 360-camera and a spherical microphone array *e.g.,* EigenMike, is able to make all parts of the surrounding environment observable.
- In this thesis, we used a stationary audio-visual sensors setup; however this is not always applicable for HRI. Often audio-visual sensors are embeeded in the robot's platform, *e.g.,* as in humanoid robot NAO. The robots has to move and navigate complex human environments. Thus, one possible direction for future works is to extend the models and algorithms developed in this thesis to accommodate the effect of the robot's physical movement in space (whether that movement is functional or expressive), *e.g.,* [Ban 17a].
- This work explores the use of sound-source localization cues and high-level visual features such as persons face locations, outputs of a visual tracker. One may add additional audio-visual cues such as head pose and head-pose tracking, in order to facilitate the detection of speech turns on the basis of people that look at each other over time. One can also incorporate richer audio features, such as the possibility to extract speech signals emitted by each participant (sound-source separation) and hence to enable not only diarization but also dialog understanding.

PUBLICATIONS

INTERNATIONAL JOURNAL PUBLICATIONS

- [Gebru 17a] I. D. Gebru, S. Ba, X. Li & R. Horaud. *Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion*. In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, 2017.
- [Gebru 16a] I. D. Gebru, X. Alameda-Pineda, F. Forbes & R. Horaud. *EM algorithms for weighted-data clustering with application to audio-visual scene analysis*. In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 12, pages 2402 – 2415, 2016.

INTERNATIONAL CONFERENCE PUBLICATIONS

- [Gebru 17b] I. D. Gebru, C. Evers, P. A. Naylor & R. Horaud. *Audio-visual tracking by density approximation in a sequential Bayesian filtering framework*. In Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017, pages 71–75. IEEE, 2017.
- [Gebru 15b] I. D. Gebru, S. Ba, G. Evangelidis & R. Horaud. *Tracking the Active Speaker Based on a Joint Audio-Visual Observation Model*. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 15–21, Santiago, Chile, 2015.
- [Gebru 15a] I. D. Gebru, S. Ba, G. Evangelidis & R. Horaud. *Audio-Visual Speech-Turn Detection and Tracking*. In The Twelfth International Conference on Latent Variable Analysis and Signal Separation, Liberec, Czech Republic, 2015
- [Gebru 14] I. D. Gebru, X. Alameda-Pineda, R. Horaud & F. Forbes. *Audio-visual speaker localization via weighted clustering*. In Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on, pages 1–6. IEEE, 2014.

OTHER ARTICLES

- [Badeig 15] F. Badeig, Q. Pelorson, S. Arias, V. Drouard, I. D. Gebru, X. Li, G. Evangelidis & R Horaud. *A distributed architecture for interacting with NAO*. In

ACM on International Conference on Multimodal Interaction (ICMI), pages 385-386, Seattle, WA, USA, November 2015.

- [Dang-Nguyen 13] D.-T. Dang-Nguyen, I. D. Gebru, V. Conotter, G. Boato & F. GB De Natale. *Counter-forensics of median filtering*. In Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on, pages 260–265. IEEE, 2013.

REFERENCES

[Ackerman 12] M. Ackerman, S. Ben-David, S. Branzei & D. Loker. *Weighted Clustering*. In Proceedings of AAAI, 2012.

[Alameda-Pineda 11] X. Alameda-Pineda, V. Khalidov, R. Horaud & F. Forbes. *Finding audio-visual events in informal social gatherings*. In ICMI,2011, pages 247–254. ACM, 2011.

[Alameda-Pineda 13] X. Alameda-Pineda, J. Sanchez-Riera, J. Wienke, V. Franc, J. Čech, K. Kulkarni, A. Deleforge & R. Horaud. *RAVEL: An annotated corpus for training robots with audiovisual abilities*. Journal on Multimodal User Interfaces, vol. 7, no. 1-2, pages 79–91, 2013.

[Alameda-Pineda 15a] X. Alameda-Pineda & R. Horaud. *Vision-Guided Robot Hearing*. The International Journal of Robotics Research, vol. 34, no. 4-5, pages 437–456, April 2015.

[Alameda-Pineda 15b] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz & N. Sebe. *Analyzing free-standing conversational groups: a multimodal approach*. In Proceedings of the 23rd ACM International Conference on Multimedia, pages 5–14, 2015.

[Anderson 79] B. D. Anderson & J. B. Moore. *Optimal filtering*. Englewood Cliffs, vol. 21, pages 22–95, 1979.

[Andersson 10] M. Andersson, S. Ntalampiras, T. Ganchev, J. Rydell, J. Ahlberg & N. Fakotakis. *Fusion of acoustic and optical sensor data for automatic fight detection in urban environments*. In Information Fusion (FUSION), 2010 13th Conference on, pages 1–8. IEEE, 2010.

[Andrews 12] J. L. Andrews & P. D. McNicholas. *Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions*. Statistics and Computing, vol. 22, no. 5, pages 1021–1029, 2012.

[Andrieu 03] C. Andrieu, N. De Freitas, A. Doucet & M. I. Jordan. *An introduction to MCMC for machine learning*. Machine learning, vol. 50, no. 1-2, pages 5–43, 2003.

[Anguera Miro 12] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland & O. Vinyals. *Speaker diarization: A review of recent research*. IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 2, pages 356–370, 2012.

[Archambeau 07] C. Archambeau & M. Verleysen. *Robust Bayesian Clustering*. Neural Networks, vol. 20, no. 1, pages 129–138, 2007.

[Arnaud 08] E. Arnaud, H. Christensen, Y.-C. Lu, J. Barker, V. Khalidov, M. Hansard, B. Holveck, H. Mathieu, R. Narasimha, E. Taillant & Others. *The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements*. In Proceedings of the 10th international conference on Multimodal interfaces, pages 109–116. ACM, 2008.

[Ba 16] S. Ba, X. Alameda-Pineda, A. Xompero & R. Horaud. *An online variational Bayesian model for multi-person tracking from cluttered scenes*. Computer Vision and Image Understanding, vol. 153, pages 64–76, 2016.

[Badeig 15] F. Badeig, Q. Pelorson, S. Arias, V. Drouard, I. Gebru, X. Li, G. Evangelidis & R. Horaud. *A distributed architecture for interacting with NAO*. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pages 385–386. ACM, 2015.

[Bae 14] S.-H. Bae & K.-J. Yoon. *Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning*. In Computer Vision and Pattern Recognition, pages 1218–1225, 2014.

[Ban 17a] Y. Ban, X. Alameda-Pineda, F. Badeig, S. Ba & R. Horaud. *Tracking a Varying Number of People with a Visually-Controlled Robotic Head*. In IEEE/RSJ International Conference on Intelligent Robots and Systems, 2017.

[Ban 17b] Y. Ban, L. Girin, X. Alameda-Pineda & R. Horaud. *Exploiting the Complementarity of Audio and Visual Data in Multi-Speaker Tracking*. In ICCV Workshop on Computer Vision for Audio-Visual Media, 2017.

[Banfield 93] J. Banfield & A. E. Raftery. *Model-based Gaussian and non-Gaussian clustering*. Biometrics, vol. 49, no. 3, pages 803–821, 1993.

[Barzelay 10] Z. Barzelay & Y. Y. Schechner. *Onsets coincidence for cross-modal analysis*. IEEE Transactions on Multimedia, vol. 12, no. 2, pages 108–120, 2010.

[Baudry 10] J. P. Baudry, E. A. Raftery, G. Celeux, K. Lo & R. Gottardo. *Combining mixture components for clustering*. Journal of Computational and Graphical Statistics, vol. 19, no. 2, 2010.

[Beal 02] M. J. Beal, H. Attias & N. Jojic. *Audio-video sensor fusion with probabilistic graphical models*. In Computer Vision—ECCV 2002, pages 736–750. Springer, 2002.

[Bernardin 08] K. Bernardin & R. Stiefelhagen. *Evaluating multiple object tracking performance: the CLEAR MOT metrics*. EURASIP Journal on Image and Video Processing, no. 1, pages 1–10, 2008.

[Besson 08] P. Besson, V. Popovici, J.-M. Vesin, J. Thiran & M. Kunt. *Extraction of audio features specific to speech production for multimodal speaker detection*. Multimedia, IEEE Transactions on, vol. 10, no. 1, pages 63–73, 2008.

[Bishop 05] C. M. Bishop & M. Svensen. *Robust Bayesian Mixture Modelling*. Neurocomputing, vol. 64, pages 235–252, 2005.

[Bishop 06] C. M. Bishop & N. M. Nasrabadi. *Pattern recognition and machine learning*. 2006.

[Blandin 12] C. Blandin, A. Ozerov & E. Vincent. *Multi-source TDOA estimation in reverberant audio using angular spectra and clustering*. Signal Processing, vol. 92, no. 8, pages 1950–1960, 2012.

[Bohus 09] D. Bohus & E. Horvitz. *Dialog in the open world: platform and applications*. In Proceedings of the 2009 international conference on Multimodal interfaces, pages 31–38. ACM, 2009.

[Bohus 10] D. Bohus & E. Horvitz. *Facilitating multiparty dialog with gaze, gesture, and speech*. In International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, page 5. ACM, 2010.

[Bohus 11] D. Bohus & E. Horvitz. *Decisions about turns in multiparty conversation: from perception to action*. In Proceedings of the 13th international conference on multimodal interfaces, pages 153–160. ACM, 2011.

[Bouguet 04] J.-Y. Bouguet. *Camera calibration toolbox for matlab*. 2004.

[Bourdev 09] L. Bourdev & J. Malik. *Poselets: Body part detectors trained using 3d human pose annotations*. In 2009 IEEE 12th International Conference on Computer Vision, pages 1365–1372. IEEE, 2009.

[Brandstein 97] M. S. Brandstein & H. F. Silverman. *A robust method for speech signal time-delay estimation in reverberant rooms*. In Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on, volume 1, pages 375–378. IEEE, 1997.

[Bregman 94] A. S. Bregman. *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[Breiman 84] L. Breiman, J. Friedman, C. J. Stone & R. A. Olshen. *Classification and Regression Trees*. CRC Press, 1984.

[Burger 02] S. Burger, V. MacLaren & H. Yu. *The ISL meeting corpus: The impact of meeting type on speech style*. In Seventh International Conference on Spoken Language Processing, 2002.

[Butz 02] T. Butz & J. Thiran. *Feature space mutual information in speech-video sequences*. In ICME, 2002.

[Carletta 05] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal & Others. *The AMI meeting corpus: A pre-announcement*. In International Workshop on Machine Learning for Multimodal Interaction, pages 28–39. Springer, 2005.

[Celeux 01] G. Celeux, S. Chrétien, F. Forbes & A. Mkhadri. *A component-wise EM algorithm for mixtures*. Journal of Computational and Graphical Statistics, vol. 10, no. 4, 2001.

[Celiktutan 17] O. Celiktutan, E. Skordos & H. Gunes. *Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement*. IEEE Transactions on Affective Computing, 2017.

[Checka 04] N. Checka, K. Wilson, M. Siracusa & T. Darrell. *Multiple person and speaker activity tracking with a particle filter*. In IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004.

[Cooke 06] M. Cooke, J. Barker, S. Cunningham & X. Shao. *An audio-visual corpus for speech perception and automatic speech recognition*. The Journal of the Acoustical Society of America, vol. 120, no. 5, pages 2421–2424, 2006.

[Cutler 00] R. Cutler & L. Davis. *Look who's talking: Speaker detection using video and audio correlation*. In Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on, volume 3, pages 1589–1592. IEEE, 2000.

[Dang-Nguyen 13] D.-T. Dang-Nguyen, I. D. Gebru, V. Conotter, G. Boato & F. G. De Natale. *Counter-forensics of median filtering*. In Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on, pages 260–265. IEEE, 2013.

[Davies 79] D. L. Davies & D. W. Bouldin. *A cluster separation measure*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, no. 2, pages 224–227, 1979.

[Deleforge 13] A. Deleforge, F. Forbes & R. Horaud. *Variational EM for binaural sound-source separation and localization*. In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, pages 76–80. IEEE, 2013.

[Deleforge 14a] A. Deleforge, V. Drouard, L. Girin & R. Horaud. *Mapping Sounds on Images Using Binaural Spectrograms*. In Proceedings of the European Signal Processing Conference, Lisbon, Portugal, 2014.

[Deleforge 14b] A. Deleforge, V. Drouard, L. Girin & R. Horaud. *Mapping Sounds on Images Using Binaural Spectrograms*. In European Signal Processing Conference, Lisbonne, Portugal, September 2014.

[Deleforge 14c] A. Deleforge, F. Forbes & R. Horaud. *Acoustic space learning for sound-source separation and localization on binaural manifolds*. International Journal of Neural Systems, vol. 0, no. February, 2014.

[Deleforge 15a] A. Deleforge, R. Horaud, Y. Y. Schechner & L. Girin. *Co-Localization of Audio Sources in Images Using Binaural Features and Locally-Linear Regression*. IEEE Transactions on Audio, Speech and Language Processing, vol. 23, no. 4, pages 718–731, April 2015.

[Deleforge 15b] A. Deleforge, R. Horaud, Y. Y. Schechner & L. Girin. *Co-Localization of Audio Sources in Images Using Binaural Features and Locally-Linear Regression*. IEEE Transactions on Audio, Speech and Language Processing, vol. 23, no. 4, pages 718–731, 2015.

[Dhillon 04] I. S. Dhillon, Y. Guan & B. Kulis. *Kernel K-means: spectral clustering and normalized cuts*. In Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining, pages 551–556. ACM, 2004.

[Dorfan 15] Y. Dorfan & S. Gannot. *Tree-based recursive expectation-maximization algorithm for localization of acoustic sources*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 10, pages 1692–1703, 2015.

[El Khoury 14] E. El Khoury, C. Sénac & P. Joly. *Audiovisual diarization of people in video content*. Multimedia tools and applications, vol. 68, no. 3, pages 747–775, 2014.

[Evers 14] C. Evers, A. H. Moore & P. A. Naylor. *Multiple source localisation in the spherical harmonic domain*. In IEEE International Workshop on Acoustic Signal Enhancement (IWAENC), 2014.

[Evers 15] C. Evers, A. H. Moore, P. A. Naylor, J. Sheaffer & B. Rafaely. *Bearing-only acoustic tracking of moving speakers for robot audition*. In International Conference on Digital Signal Processing (DSP), 2015.

[Feldman 12] D. Feldman & L. Schulman. *Data reduction for weighted and outlier-resistant clustering*. In Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1343–1354. SIAM, 2012.

[Ferrari 08] V. Ferrari, M. Marin-Jimenez & A. Zisserman. *Progressive search space reduction for human pose estimation*. In Computer Vision and Pattern Recognition, pages 1–8, 2008.

[Figueiredo 02] M. A. T. Figueiredo & A. K. Jain. *Unsupervised learning of finite mixture models*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pages 381–396, 2002.

[Fiscus 05] J. G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot & C. Laprun. *The rich transcription 2005 spring meeting recognition evaluation*. In International Workshop on Machine Learning for Multimodal Interaction, pages 369–389. Springer, 2005.

[Fiscus 06] J. G. Fiscus, J. Ajot, M. Michel & J. S. Garofolo. *The rich transcription 2006 spring meeting recognition evaluation*. In International Workshop on Machine Learning for Multimodal Interaction, pages 309–322. Springer, 2006.

[Fisher III 00] J. W. Fisher III, T. Darrell, W. T. Freeman & P. A. Viola. *Learning joint statistical models for audio-visual fusion and segregation*. In NIPS, pages 772–778, 2000.

[Fisher 04] J. W. Fisher & T. Darrell. *Speaker association with signal-level audiovisual fusion*. IEEE TMM, vol. 6, no. 3, pages 406–413, 2004.

[Forbes 10] F. Forbes, S. Doyle, D. Garcia-Lorenzo, C. Barillot, M. Dojat*et al.* *A weighted multi-sequence Markov model for brain lesion segmentation*. In JMLR Workshop and Conference Proceedings Volume 9: AISTATS 2010, volume 9, pages 225–232, 2010.

[Forbes 14] F. Forbes & D. Wraith. *A new family of multivariate heavy-tailed distributions with variable marginal amounts of tailweight: application to robust clustering*. Statistics and Computing, vol. 24, no. 6, pages 971–984, 2014.

[Frey 91] P. W. Frey & D. J. Slate. *Letter recognition using Holland-style adaptive classifiers*. Machine Learning, vol. 6, no. 2, pages 161–182, 1991.

[Garau 10] G. Garau, A. Dielmann & H. Bourlard. *Audio-visual synchronisation for speaker diarisation.* In INTERSPEECH, pages 2654–2657, 2010.

[Garofolo 93] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus & D. S. Pallett. *DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM*. NASA STI/Recon technical report n, vol. 93, 1993.

[Garofolo 04] J. S. Garofolo, C. Laprun, M. Michel, V. M. Stanford & E. Tabassi. *The NIST Meeting Room Pilot Corpus.* In LREC, 2004.

[Gatica-Perez 07] D. Gatica-Perez, G. Lathoud, J.-M. Odobez & I. McCowan. *Audiovisual probabilistic tracking of multiple speakers in meetings*. IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 2, pages 601–616, 2007.

[Gatica-Perez 09] D. Gatica-Perez. *Automatic nonverbal analysis of social interaction in small groups: A review*. Image and Vision Computing, vol. 27, no. 12, pages 1775–1787, 2009.

[Gebru 14] I. D. Gebru, X. Alameda-Pineda, R. Horaud & F. Forbes. *Audio-visual speaker localization via weighted clustering*. In Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on, pages 1–6. IEEE, 2014.

[Gebru 15a] I. D. Gebru, S. Ba, G. Evangelidis & R. Horaud. *Audio-Visual Speech-Turn Detection and Tracking*. In The Twelfth International Conference on Latent Variable Analysis and Signal Separation, Liberec, Czech Republic, 2015.

[Gebru 15b] I. D. Gebru, S. Ba, G. Evangelidis & R. Horaud. *Tracking the Active Speaker Based on a Joint Audio-Visual Observation Model*. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 15–21, Santiago, Chile, 2015.

[Gebru 16a] I. D. Gebru, X. Alameda-Pineda, F. Forbes & R. Horaud. *EM algorithms for weighted-data clustering with application to audio-visual scene analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 12, pages 2402 – 2415, 2016.

[Gebru 16b] I. Gebru, X. Alameda-Pineda, F. Forbes & R. Horaud. *{EM} Algorithms for Weighted-Data Clustering with Application to Audio-Visual Scene Analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016.

[Gebru 17a] I. D. Gebru, S. Ba, X. Li & R. Horaud. *Audio-Visual Speaker Diarization Based on Spatiotemporal Bayesian Fusion*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, 2017.

[Gebru 17b] I. D. Gebru, C. Evers, P. A. Naylor & R. Horaud. *Audio-visual tracking by density approximation in a sequential Bayesian filtering framework*. In Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017, pages 71–75. IEEE, 2017.

[Görür 10] D. Görür & C. E. Rasmussen. *Dirichlet process gaussian mixture models: Choice of the base distribution*. Journal of Computer Science and Technology, vol. 25, no. 4, pages 653–664, 2010.

[Gurban 06] M. Gurban & J.-P. Thiran. *Multimodal speaker localization in a probabilistic framework*. In Signal Processing Conference, 2006 14th European, pages 1–5. IEEE, 2006.

[Haykin 01] S. S. Haykin*et al.* *Kalman filtering and neural networks*. Wiley Online Library, 2001.

[Hazen 04] T. J. Hazen, K. Saenko, C.-H. La & J. R. Glass. *A segment-based audio-visual speech recognizer: Data collection, development, and initial experiments*. In Proceedings of the 6th international conference on Multimodal interfaces, pages 235–242. ACM, 2004.

[Hennig 10] C. Hennig. *Methods for merging Gaussian mixture components*. Advances in Data Analysis and Classification, vol. 4, no. 1, pages 3–34, 2010.

[Hershey 00] J. Hershey & J. Movellan. *Audio-vision: Using audio-visual synchrony to locate sounds*. In Advances in Neural Information Processing Systems, pages 813–819, 2000.

[Hoffman 04] J. R. Hoffman & R. P. Mahler. *Multitarget miss distance via optimal assignment*. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, vol. 34, no. 3, pages 327–336, 2004.

[Itti 98] L. Itti, C. Koch, E. Niebur & Others. *A model of saliency-based visual attention for rapid scene analysis*. IEEE Transactions on pattern analysis and machine intelligence, vol. 20, no. 11, pages 1254–1259, 1998.

[Janin 03] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke*et al. The ICSI meeting corpus*. In Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on, volume 1, pages I–I. IEEE, 2003.

[Jayagopi 13] D. B. Jayagopi, S. Sheiki, D. Klotz, J. Wienke, J.-M. Odobez, S. Wrede, V. Khalidov, L. Nyugen, B. Wrede & D. Gatica-Perez. *The vernissage corpus: A conversational human-robot-interaction dataset*. In Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction, pages 149–150. IEEE Press, 2013.

[Johansson 14] M. Johansson, G. Skantze & J. Gustafson. *Comparison of Human-Human and Human-Robot Turn-Taking Behaviour in Multiparty Situated Interaction*. In Proceedings of the 2014 workshop on Understanding and Modeling Multiparty, Multimodal Interactions, pages 21–26. ACM, 2014.

[Julier 96] S. J. Julier & J. K. Uhlmann. *A general method for approximating nonlinear transformations of probability distributions*. Rapport technique, Robotics Research Group, Department of Engineering Science, University of Oxford, 1996.

[Julier 97] S. J. Julier & J. K. Uhlmann. *New extension of the Kalman filter to nonlinear systems*. pages 182–193. International Society for Optics and Photonics, 1997.

[Kächele 14] M. Kächele, S. Meudt, A. Schwarz & F. Schwenker. *Audio-Visual User Identification in HCI Scenarios*. In IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction, pages 113–122. Springer, 2014.

[Kapsouras 16] I. Kapsouras, A. Tefas, N. Nikolaidis, G. Peeters, L. Benaroya & I. Pitas. *Multimodal speaker clustering in full length movies*. Multimedia Tools and Applications, 2016.

[Khalidov 08a] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud & R. Horaud. *Audio-Visual clustering for 3D speaker localization*. In International Workshop on Machine Learning for Multimodal Interaction, pages 86–97. Springer, 2008.

[Khalidov 08b] V. Khalidov, F. Forbes, M. Hansard, E. Arnaud & R. Horaud. *Detection and localization of 3d audio-visual objects using unsupervised clustering*. In ICMI, pages 217–224. ACM, 2008.

[Khalidov 11a] V. Khalidov, F. Forbes & R. Horaud. *Conjugate Mixture Models for Clustering Multimodal Data*. Neural Computation, vol. 23, no. 2, pages 517–557, 2011.

[Khalidov 11b] V. Khalidov, F. Forbes & R. Horaud. *Conjugate mixture models for clustering multimodal data*. Neural Computation, vol. 23, no. 2, pages 517–557, 2011.

[Khalidov 13] V. Khalidov, F. Forbes & R. Horaud. *Alignment of Binocular-Binaural Data Using a Moving Audio-Visual Target*. In IEEE Workshop on Multimedia Signal Processing, Pula, Italy, 2013.

[Kidron 05] E. Kidron, Y. Y. Schechner & M. Elad. *Pixels that sound*. In Computer Vision and Pattern Recognition, volume 1, pages 88–95, 2005.

[Kidron 07] E. Kidron, Y. Schechner & M. Elad. *Cross-modal localization via sparsity*. Signal Processing, IEEE Transactions on, vol. 55, no. 4, pages 1390–1404, 2007.

[Kilic 15a] V. Kilic, M. Barnard, W. Wang & J. Kittler. *Audio Assisted Robust Visual Tracking With Adaptive Particle Filtering*. IEEE Transactions on Multimedia, vol. 17, no. 2, pages 186–200, 2015.

[Kılıç 15b] V. Kılıç, M. Barnard, W. Wang & J. Kittler. *Audio assisted robust visual tracking with adaptive particle filtering*. IEEE Transactions on Multimedia, vol. 17, no. 2, pages 186–200, 2015.

[Kotz 04] S. Kotz & S. Nadarajah. *Multivariate t Distributions and their Applications*. Cambridge University Press, 2004.

[Kuhn 55] H. W. Kuhn. *The Hungarian method for the assignment problem*. Naval research logistics quarterly, vol. 2, no. 1-2, pages 83–97, 1955.

[Lang 03] S. Lang, M. Kleinehagenbrock, S. Hohenner, J. Fritsch, G. A. Fink & G. Sagerer. *Providing the basis for human-robot-interaction: A multi-modal attention system for a mobile robot*. In International conference on Multimodal interfaces. ACM, 2003.

[Lathoud 04] G. Lathoud, J.-M. Odobez & D. Gatica-Perez. *AV16. 3: an audio-visual corpus for speaker localization and tracking*. In Machine Learning for Multimodal Interaction, pages 182–195. Springer, 2004.

[Lathoud 05] G. Lathoud, J.-m. Odobez & D. Gatica-perez. *AV16 . 3 : An Audio-Visual Corpus for Speaker Localization and Tracking*. pages 182–195, 2005.

[LeCun 98] Y. LeCun, L. Bottou, Y. Bengio & P. Haffner. *Gradient-based learning applied to document recognition*. Proceedings of the IEEE, vol. 86, no. 11, pages 2278–2324, 1998.

[Lee 14] S. Lee & G. J. McLachlan. *Finite mixtures of multivariate skew t-distributions: some recent and new results*. Statistics and Computing, vol. 24, no. 2, pages 181–202, 2014.

[Li 15a] X. Li, L. Girin, R. Horaud & S. Gannot. *Estimation of Relative Transfer Function in the Presence of Stationary Noise Based on Segmental Power Spectral Density Matrix Subtraction*. In IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia, 2015.

[Li 15b] X. Li, R. Horaud, L. Girin & S. Gannot. *Local Relative Transfer Function for Sound Source Localization*. In European Signal Processing Conference, Nice, France, 2015.

[Li 16] X. Li, L. Girin, S. Gannot & R. Horaud. *Non-stationary noise power spectral density estimation based on regional statistics*. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 181–185. IEEE, 2016.

[Long 06] B. Long, Z. M. Zhang, X. Wu & P. S. Yu. *Spectral clustering for multi-type relational data*. In Proceedings of the 23rd International Conference on Machine learning, pages 585–592. ACM, 2006.

[Mandel 10] M. I. Mandel, R. J. Weiss & D. P. W. Ellis. *Model-based expectation-maximization source separation and localization*. IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 2, pages 382–394, 2010.

[McCool 12] C. McCool, S. Marcel, A. Hadid, M. Pietikäinen, P. Matejka, J. Cernockỳ, N. Poh, J. Kittler, A. Larcher, C. Levy*et al. Bi-modal person recognition on a mobile phone: using mobile phone data*. In Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on, pages 635–640. IEEE, 2012.

[McLachlan 00a] G. McLachlan & D. Peel. *Finite Mixture Models*. Wiley, 2000.

[McLachlan 00b] G. McLachlan & D. Peel. *Robust Mixture Modelling Using the t Distribution*. Statistics and Computing, vol. 10, no. 4, pages 339–348, 2000.

[Melnykov 14] V. Melnykov. *Merging mixture components for clustering through pairwise overlap*. Journal of Computational and Graphical Statistics, 2014.

[Minotto 15] V. P. Minotto, C. R. Jung & B. Lee. *Multimodal On-line Speaker Diarization using Sensor Fusion through {SVM}*. IEEE Transactions on Multimedia, vol. 17, no. 10, pages 1694–1705, 2015.

[Mohammad 08] Y. Mohammad, Y. Xu, K. Matsumura & T. Nishida. *The H3R Explanation Corpus human-human and base human-robot interaction dataset*. In Intelligent Sensors, Sensor Networks and Information Processing, 2008. ISSNIP 2008. International Conference on, pages 201–206. IEEE, 2008.

[Naqvi 10a] S. Naqvi, M. Yu & J. Chambers. *A Multimodal Approach to Blind Source Separation of Moving Sources*. IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 5, pages 895 –910, 2010.

[Naqvi 10b] S. M. Naqvi, M. Yu & J. A. Chambers. *A multimodal approach to blind source separation of moving sources*. IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 5, pages 895–910, 2010.

[Nock 03] H. J. Nock, G. Iyengar & C. Neti. *Speaker localisation using audio-visual synchrony: An empirical study*. In International conference on image and video retrieval, pages 488–499. Springer, 2003.

[Noulas 07] A. Noulas & B. J. A. Krose. *On-line multi-modal speaker diarization*. In Proceedings of the 9th international conference on Multimodal interfaces, pages 350–357. ACM, 2007.

[Noulas 12] A. Noulas, G. Englebienne & B. Krose. *Multimodal speaker diarization*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no. 1, pages 79–93, 2012.

[Otsuka 07] K. Otsuka, H. Sawada & J. Yamato. *Automatic inference of cross-modal nonverbal interactions in multiparty conversations: who responds to whom, when, and how? from gaze, head gestures, and utterances*. In Proceedings of the 9th international conference on Multimodal interfaces, pages 255–262. ACM, 2007.

[Patterson 02] E. K. Patterson, S. Gurbuz, Z. Tufekci & J. N. Gowdy. *CUAVE: A new audio-visual database for multimodal human-computer interface research*. In Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on, volume 2, pages II–2017. IEEE, 2002.

[Poggio 89] T. Poggio & F. Girosi. *A theory of networks for approximation and learning*. Rapport technique, DTIC Document, 1989.

[Potamianos 03] G. Potamianos, C. Neti, G. Gravier, A. Garg & A. W. Senior. *Recent advances in the automatic recognition of audiovisual speech*. Proceedings of the IEEE, vol. 91, no. 9, pages 1306–1326, 2003.

[Rasmussen 99] C. E. Rasmussen. *The infinite Gaussian mixture model.* In NIPS, volume 12, pages 554–560, 1999.

[Rivet 07] B. Rivet, L. Girin & C. Jutten. *Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures*. IEEE transactions on audio, speech, and language processing, vol. 15, no. 1, pages 96–108, 2007.

[Sanchez-Riera 12] J. Sanchez-Riera, X. Alameda-Pineda, J. Wienke, A. Deleforge, S. Arias, J. Čech, S. Wrede & R. Horaud. *Online multimodal speaker detection for humanoid robots*. In IEEE International Conference on Humanoid Robots, 2012.

[Sarafianos 16] N. Sarafianos, T. Giannakopoulos & S. Petridis. *Audio-visual speaker diarization using fisher linear semi-discriminant analysis*. Multimedia Tools and Applications, vol. 75, no. 1, pages 115–130, 2016.

[Sargin 07] M. E. Sargin, Y. Yemez, E. Erzin & M. A. Tekalp. *Audiovisual synchronization and fusion using canonical correlation analysis*. IEEE Transactions on Multimedia, vol. 9, no. 7, pages 1396–1403, 2007.

[Schuhmacher 08] D. Schuhmacher, B.-T. Vo & B.-N. Vo. *A consistent metric for performance evaluation of multi-object filters*. IEEE Transactions on Signal Processing, vol. 56, no. 8, pages 3447–3457, 2008.

[Schwarz 78] G. Schwarz*et al.* *Estimating the dimension of a model*. The Annals of Statistics, vol. 6, no. 2, pages 461–464, 1978.

[Sheather 91] S. J. Sheather & M. C. Jones. *A reliable data-based bandwidth selection method for kernel density estimation*. Journal of the Royal Statistical Society. Series B (Methodological), pages 683–690, 1991.

[Shi 00] J. Shi & J. Malik. *Normalized cuts and image segmentation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pages 888–905, 2000.

[Shivappa 10] S. T. Shivappa, M. M. Trivedi & B. D. Rao. *Audiovisual information fusion in human–computer interfaces and intelligent environments: A survey*. Proceedings of the IEEE, vol. 98, no. 10, pages 1692–1715, 2010.

[Siracusa 07] M. R. Siracusa & J. W. Fisher. *Dynamic dependency tests for audio-visual speaker association*. In 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, volume 2, pages II—-457. IEEE, 2007.

[Skantze 14] G. Skantze, A. Hjalmarsson & C. Oertel. *Turn-taking, feedback and joint attention in situated human–robot interaction*. Speech Communication, vol. 65, pages 50–66, 2014.

[Sohn 99] J. Sohn, N. S. Kim & W. Sung. *A statistical model-based voice activity detection*. IEEE Signal Processing Letters, vol. 6, no. 1, pages 1–3, 1999.

[Stiefelhagen 06] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa & P. Soundararajan. *The CLEAR 2006 evaluation*. In International Evaluation Workshop on Classification of Events, Activities and Relationships, pages 1–44. Springer, 2006.

[Stiefelhagen 08] R. Stiefelhagen, R. Bowers & J. Fiscus. *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers*, volume 4625. Springer, 2008.

[Street 93] W. Street, W. Wolberg & O. Mangasarian. *Nuclear feature extraction for breast tumor diagnosis*. In IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology, pages 861–870. International Society for Optics and Photonics, 1993.

[Sun 10] J. Sun, A. Kabán & J. M. Garibaldi. *Robust mixture clustering using Pearson type VII distribution*. Pattern Recognition Letters, vol. 31, no. 16, pages 2447–2454, 2010.

[Talantzis 08] F. Talantzis, A. Pnevmatikakis & A. G. Constantinides. *Audio–visual active speaker tracking in cluttered indoors environments*. Cybernetics, IEEE Transactions on, vol. 38, no. 3, pages 799–807, 2008.

[Tseng 07] G. Tseng. *Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data*. Bioinformatics, vol. 23, no. 17, pages 2247–2255, 2007.

[Van Der Merwe 00] R. Van Der Merwe, A. Doucet, N. De Freitas & E. Wan. *The unscented particle filter*. In Advances In Neural Information Processing Systems, pages 584–590, 2000.

[Van Veen 88] B. D. Van Veen & K. M. Buckley. *Beamforming: A versatile approach to spatial filtering*. IEEE ASSP Magazine, vol. 5, no. 2, pages 4–24, 1988.

[Vijayasenan 12] D. Vijayasenan & F. Valente. *DiarTk: An Open Source Toolkit for Research in Multistream Speaker Diarization and its Application to Meetings Recordings.* In INTERSPEECH, pages 2170–2173, 2012.

[Vinciarelli 12] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico & M. Schroeder. *Bridging the gap between social animal and unsocial machine: A survey of social signal processing*. IEEE Transactions on Affective Computing, vol. 3, no. 1, pages 69–87, 2012.

[Viola 04] P. Viola & M. J. Jones. *Robust real-time face detection*. International Journal of Computer Vision, vol. 57, no. 2, pages 137–154, 2004.

[Wei 12] X. Wei & C. Li. *The infinite Student's t-mixture for robust modeling*. Signal Processing, vol. 92, no. 1, pages 224–234, 2012.

[Wooters 08] C. Wooters & M. Huijbregts. *The ICSI RT07s speaker diarization system*. In Multimodal Technologies for Perception of Humans, pages 509–519. Springer, 2008.

[Xi 04] W. Xi, B. Zhang, Z. Chen, Y. Lu, S. Yan, W.-Y. Ma & E. A. Fox. *Link fusion: a unified link analysis framework for multi-type interrelated data objects*. In Proceedings, WWW2004, pages 319–327. ACM, 2004.

[Yan 13] Y. Yan, E. Ricci, R. Subramanian, O. Lanz & N. Sebe. *No matter where you are: Flexible graph-guided multi-task learning for multi-view head pose classification under target motion*. In IEEE International Conference on Computer Vision, pages 1177–1184, 2013.

[Yehia 98] H. Yehia, P. Rubin & E. Vatikiotis-Bateson. *Quantitative association of vocal-tract and facial behavior*. Speech Communication, vol. 26, no. 1, pages 23–43, 1998.

[Yerebakan 14] H. Z. Yerebakan, B. Rajwa & M. Dundar. *The Infinite Mixture of Infinite Gaussian Mixtures*. In Advances in Neural Information Processing Systems, pages 28–36, 2014.

[Zancanaro 06] M. Zancanaro, B. Lepri & F. Pianesi. *Automatic detection of group functional roles in face to face interactions*. In Proceedings of the 8th international conference on Multimodal interfaces, pages 28–34. ACM, 2006.

[Zhao 02] Y. Zhao & G. Karypis. *Evaluation of hierarchical clustering algorithms for document datasets*. In Proceedings of the Eleventh International Conference on Information and Knowledge Management, pages 515–524. ACM, 2002.

[Zhao 12] X. Zhao, N. Evans & J.-L. Dugelay. *Co-lda: A semi-supervised approach to audio-visual person recognition*. In Multimedia and

Expo (ICME), 2012 IEEE International Conference on, pages 356–361. IEEE, 2012.

[Zhou 08] H. Zhou, M. Taj & A. Cavallaro. *Target detection and tracking with heterogeneous sensors*. IEEE Journal of Selected Topics in Signal Processing, vol. 2, no. 4, pages 503–513, 2008.

[Zhu 12] X. Zhu & D. Ramanan. *Face detection, pose estimation, and landmark localization in the wild*. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2879–2886. IEEE, 2012.

[Zotkin 02] D. Zotkin, R. Duraiswami & L. Davis. *Joint audio-visual tracking using particle filters*. EURASIP Journal on Applied Signal Processing, vol. 2002, no. 1, pages 1154–1164, 2002.